The Darwinian Returns to Scale

David Rezza Baqaee
UCLA

Emmanuel Farhi[*]
Harvard

November 18, 2019

Abstract

How does an increase in the size of a market, say due to fertility, immigration, or trade, affect welfare and real GDP? We study this question in the context of a model with heterogeneous firms, monopolistic competition, and increasing returns. An increase in the size of the market improves technical efficiency by inducing entry since fixed costs can be spread over more customers. More interestingly, an increase in market size also toughens competition, reduces markups, and triggers Darwinian reallocations across firms: large firms expand, and small firms shrink or exit. Our analysis shows that changes in allocative efficiency, due to reallocations across heterogenous firms, are quantitatively much more important than the aforementioned change in technical efficiency. Using firm-level information, we non-parametrically identify residual demand curves, and using these estimates, quantify our theoretical results. We find that somewhere between 70 to 90% of the welfare effects of a change in population are due to changes in allocative efficiency. Furthermore, these reallocation effects are not driven by the oft-emphasized pro-competitive (markup-reducing) effects of market size.

[*] Emails: baqaee@econ.ucla.edu, efarhi@harvard.edu. We thank Cédric Duprez and Oleg Itskhoki for sharing their data. We thank Maria Voronina and Sihwan Yang for outstanding research assistance.

1 Introduction

Increasing returns to scale provide an incentive for trade and a mechanism for growth. In many models of trade and growth, a key question is how welfare and output respond to changes in the size of the market. For growth, the size of the market typically depends on the size of the population, where a greater population allows the fixed costs of creating ideas to be spread across a larger group, thereby raising the standard of living. For trade, as markets become integrated, it becomes viable to produce a greater variety of goods to service the larger market, once again providing gains from market integration and trade.

When efficient, a decentralized economy behaves (at least locally) like a planning problem. So, we can understand how a change in population affects aggregate welfare by studying only the technological aspects of the problem (the aspects of the problem relevant to a social planner): a bigger population allows the planner to better exploit scale economies, and the strength of these scale effects determines how beneficial population growth is.

However, scale economies are oftentimes linked to market power and, hence, to inefficiency. This is important because if the world is inefficient, then changes in market size fundamentally entangle technical and allocative efficiency together. Increases in market size intensify competition amongst firms, and these Darwinian forces trigger reallocations among a multitude of margins. If the economy were efficient, the envelope theorem would guarantee that these reallocations cancel out to a first-order. However, when the economy is initially inefficient, these reallocation effects may amplify or mitigate the pure technological effects of the shock.

This paper is a study of this problem. We characterize how welfare and real GDP respond to changes in population size in a model with monopolistic competition and flexible Kimball (1995) demand. We characterize changes in technical and allocative efficiency separately, and decompose changes in allocative efficiency into the different margins of adjustment.

In response to a shock, there are three margins along which resources can be reallocated: (1) the share of resources allocated to fixed costs versus variable costs, (2) the share of variable costs amongst existing producers (how much each type of firm produces in equilibrium), and (3) the allocation of fixed costs across firm types (how many firms of each type operate in equilibrium). The decentralized equilibrium may be inefficient along each margin: (1) entry inefficiency — the total amount of resources dedicated to entry relative to variable production could be inefficient, (2) relative production inefficiency — the allocation of resources for variable production across existing firms is inefficient, or (3) selection inefficiency — resources on fixed costs are spent on the wrong type of firms, we may have too many small firms or too many large firms in equilibrium. This paper characterizes how each of these margins moves in the decentralized equilibrium.

Our results are non-parametric and allow the residual demand curve facing individual producers to have any downward-sloping shape. We use cross-sectional firm-level information on markup elasticities (from Amiti et al., 2019) to non-parametrically identify household preferences, and given these estimates, we quantify how welfare and GDP change in response to shocks. We also decompose the overall effect into movements along each of the aforementioned margins. This procedure is especially useful since there is no consensus on a parametric functional form for representing preferences. The specification of preferences is crucial, and our data-driven approach allows us the freedom to match the data in terms of both pass-throughs and sales shares, whereas typical parametric specifications of Kimall preferences, for instance CES or Klenow and Willis (2016), have counterfactual properties.[1]

Furthermore, by separately characterizing the behavior of welfare and GDP, we clarify some potentially confusing issues. It is well-known that when the set of goods can change due to entry and exit, real GDP and welfare may not be the same. We provide explicit formulas for welfare and GDP, and show that the two do not even need to move in the same direction. So, intuitions that apply to welfare cannot naively be used to understand the behavior of real GDP (or vice versa). In particular, we show how common intuitions can be wrong. For example, increases in the productivity cutoff, as in the Hopenhayn (1992) or Melitz (2003) model, do *not* imply that aggregate productivity or welfare must increase. As another example, the fact that firms at the cutoff are small in terms of sales does *not* imply that the selection margin is negligible.

When we apply our method to the data, we find that the vast majority of the positive welfare effects of an increase in market size are due to changes in allocative efficiency (reallocation effects). However, these beneficial reallocation effects are not due to the much-discussed pro-competitive effects of market size. In fact, in our benchmark calibration, we find that the pro-competitive effects of market size are harmful! In the following paragraphs, we give a brief account of where the positive effect comes from.

[1] With monopolistic competiton, isoelastic/CES preferences imply that the pass-through of marginal cost into markups must be constant and equal to one, whereas Klenow and Willis (2016) preferences imply that pass-throughs go to zero too quickly (equivalently, markups increase too rapidly) for very productive firms, so that without demand shifters, it becomes impossible to match the fat right-tail of the firm size distribution (see Amiti et al., 2019, for a discussion of these issues).

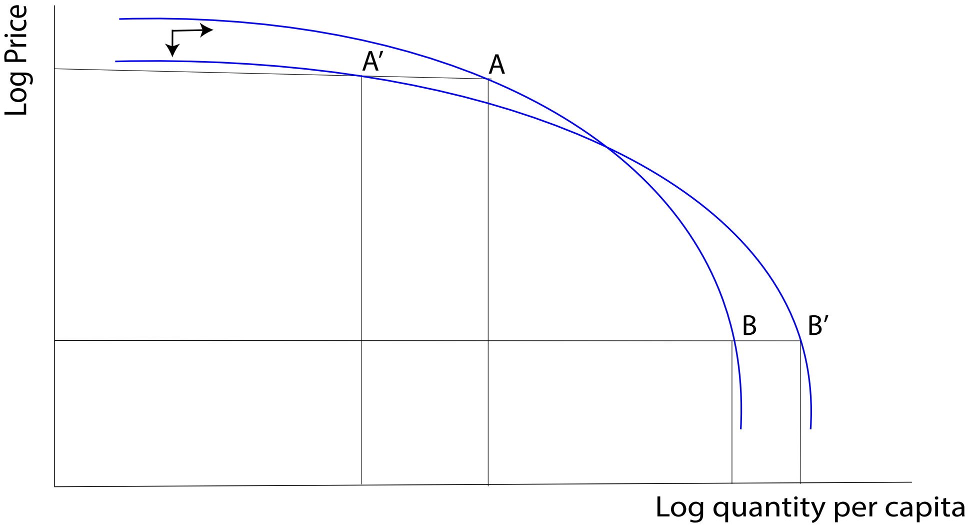


Figure 1: Illustration of the reallocation effect due to increased entry (holding fixed markups and the selection cutoff).

As population increases, more firms enter (since the fixed costs of entry are now spread over a larger market), but surprisingly, the beneficial effect is *not* due to the fact that entry is good per se. Instead, the intuition depends crucially on how this additional entry reallocates resources amongst firms. If there is heterogeneity in firm sizes, as more firms enter, holding fixed markups and the productivity cutoff, resources will be reallocated across producers. This is because larger firms face more inelastic demand curves, so as new firms enter, competition from new entrants disproportionately affects the sales of (high elasticity) small firms relative to (low elasticity) large firms.

To see the intuition, consider the residual (per capita) demand curve for a product

$$\frac{p}{P} = \Upsilon'(\frac{y}{Y}),$$

where p and y are the price and quantity of the product, P and Y are aggregate price and utility/welfare, and Υ' is some downward sloping function. Imagine the residual demand curve in log-log space, illustrated in Figure 1. An increase in entry has two effects: it shifts the curve down, because the aggregate price index P falls in response to entry, and it shifts the curve to the right (because utility rises). This means that demand per capita falls by more for the relatively elastic firms (A to A') as compared to the inelastic ones (B to B'). In fact, the demand for the highly elastic firms can increase if the rightward shift is large enough (this is the case illustrated in the figure).

So, the elastic firms lose more demand than the inelastic firms. The new entrants into the market are, in steady-state, copies of the already existing firms, so overall, resources are shifted away from the smaller firms facing more elastic demand and towards larger firms with more inelastic demand.

Crucially, it is the large firms (facing inelastic demand) that are inefficiently too small to begin with (because they have relatively high markups). Therefore, entry reallocates resources across firm types — away from low-markup small firms and towards high-markup large firms. This is overwhelmingly the largest positive force in the model, and is the source of the "positive" reallocation. Note that it relies critically on the fact that the firm size distribution be non-degenerate. If the firm-size distribution was degenerate (homogeneous firms), then this effect would disappear. Furthermore, this effect does *not* rely on whether or not we have too much or too little entry at the initial allocation.

Of course, in equilibrium both the markups and the selection cutoff will change. We find that entry intensifies competition, driving up the marginal/cutoff productivity level and driving down markups. However, quantitatively, we find both of these effects end up making things worse not better. First, the fact that the minimum productivity threshold goes up reduces welfare because (empirically) the marginal firm produces more infra-marginal value for the consumer than the average firm, and hence, driving those firms out of business hurts the consumer. Intuitively, due to the presence of fixed costs, when a firm enters, it enters at some nonzero size and the firm's value to the consumer is given by the area under the demand curve. However, the firm's decision to enter is determined by the markups it can charge, not by the area under the demand curve. In particular, we find that the selection cutoff in equilibrium is too tough, so that increasing it further makes the household worse off.

Next, the reduction in markups, also known as the pro-competitive effects of market size, are also deleterious. In response to increased entry, all firms cut their markups since demand per capita has fallen to accommodate the new entrants. The high-markup firms cut their markups by more than the low-markup firms. However, the low-markup firms face demand curves which are much more elastic. The overall reallocation effect therefore depends on a race between the reduction in markup and the elasticity of demand. Quantitatively, we find that the elasticity effect dominates, and so the reductions in markups reallocate resources away from relatively high-markup firms. However, since these firms were too small to begin with (since they had higher markups), this reallocation effect is harmful. The overall effect in equilibrium combines all these effects.

The structure of the rest of the paper is as follows. Section 2 sets up the model and defines the equilibrium, Section 3 describes the solution strategy and discusses efficiency, Section 4 analyzes the case where all firms are symmetric (homogeneous firms), while Section 5 considers the case with heterogeneous firms, Section 6 characterizes the distance from the efficient frontier, Section 7 describes how to empirically back out preferences from the data and contains our empirical application, and Section 8 concludes.

Related Literature. This paper builds on the vast literature on entry and monopolistic competition, with its origin in the works of Chamberlin (1933) and Robinson (1933). We base our analysis on the foundation of a representative consumer with a taste for variety, following Spence (1976) and Dixit and Stiglitz (1977).

Initially, the theoretical analysis of monopolistic competition was undertaken under the assumption that firms are symmetric, for example Krugman (1979), Mankiw and Whinston (1986), Vives (1999), or Venables (1985). The heterogeneous firms case has been studied by Melitz (2003) when efficient, and by Zhelobodko et al. (2012) and Dhingra and Morrow (2019) when inefficient, building on the symmetric firm analysis of Krugman (1979). Since the model in Melitz (2003) has an efficient equilibrium, the envelope theorem implies that reallocations, for example the movement in the cutoff, have no direct effect on welfare to a firs-order. Dhingra and Morrow (2019) is the closest paper to ours, since they also study inefficient models, but their focus is primarily on comparing the decentralized equilibrium to first-best, and providing qualitative conditions under which the effect of market expansion can be signed.

This paper contributes to the literature in several ways: first we provide comparative statics without imposing strong conditions on preferences. We decompose the overall effect into technical and allocative efficiency, and decompose allocative efficiency into adjustments along different margins. Second, we analyze real GDP as well as welfare, clarifying the similarities and differences between the object we typically measure (real GDP) and the one we care about (welfare). This is important since empirical studies sometimes conflate the two, and use intuitions that apply to welfare when studying real GDP (or vice versa). Third, we provide an empirical strategy for backing out household preferences from the data, allowing us to quantify our comparative static results. Finally, we provide analytical formulas for the economy's distance from the Pareto-efficient frontier, thereby explicitly linking our results to the vast literature on cross-sectional misallocation (e.g. Restuccia and Rogerson, 2008 and Hsieh and Klenow, 2009) and the welfare costs of markups (e.g. Baqaee and Farhi, 2017a, Edmond et al., 2018 or Bilbiie et al., 2019). Our decomposition of aggregate changes into pure changes in technology and changes in allocative efficiency extends the definition in Baqaee and Farhi (2017a) to economies with explicit entry and exit.

2 Model

In this section, we specify the model and describe the equilibrium.

2.1 Set Up

Households. There is a population of L identical consumers. Each consumer supplies one unit of labor and consumes different varieties of final goods indexed by $\omega \in \mathbb{R}^{+}$. Consumers have homothetic Kimball (1995) preferences, with utility Y defined implicitly in units of consumption by

$$\int_0^\infty \Upsilon(\frac{y_\omega}{Y})d\omega = 1, \tag{1}$$

where y_ω is the consumption of variety ω and Υ is an increasing and concave function in units of utils with $\Upsilon(0) = 0$.

Consumers maximize their utility Y subject to the following budget constraint

$$\int_0^\infty p_\omega y_\omega d\omega = w,$$

where p_ω is the price of variety ω and w is the wage, anticipating the result that in equilibrium, there will be no profits because of free entry. The *demand index* $\bar{\delta}$ is defined as

$$1/\bar{\delta} = \int_0^\infty \Upsilon'(y_\omega/Y)(y_\omega/Y)d\omega.$$

The per-capita inverse-demand curve for each individual variety is given by

$$p = w\frac{\bar{\delta}}{Y}\Upsilon'(\frac{y}{Y}). \tag{2}$$

Equation (2) demonstrates the appeal of Kimball preferences — by choosing Υ', we can generate demand curves of any desired (downward-sloping) shape. Equation (2) can be thought of as a relative demand curve for y/Y as a function of the relative price $(p(y)/w)/(\bar{\delta}/Y)$. In other words, $\bar{\delta}/Y$ acts like an aggregate price index for substitution. For this reason, we refer to $\bar{\delta}/Y$ as the aggregate "price index." However, we warn the reader that $\bar{\delta}/Y$ is *not* the ideal price index for the representative consumer, nor does it correspond to how price indices are usually measured in the data. In fact, let P^Y be the ideal price index and normalize the nominal wage $w = 1$, then $\bar{\delta}/Y = \bar{\delta}P^Y$.[2]

For concreteness, consider the CES special case $\Upsilon(x) = x^{\frac{\sigma-1}{\sigma}}$ where σ is the elasticity of substitution. In this case, the price index for substitution is proportional to the ideal price index $\bar{\delta}/Y = (\sigma - 1)/\sigma P^Y$. In general however, $\bar{\delta}$ is not a constant, and so $\bar{\delta}/Y$ does not move one-for-one with the ideal price index. Going forward, we refer to $\bar{\delta}/Y = \bar{\delta}P^Y$ as "the" price index without further qualification, despite the fact that it is not the same as the ideal price index.

Monopolistic competition models with Kimball demand are parsimonious in the sense that firms compete against each other via this aggregate index. Relative demand for a good is determined by the ratio of the good's price relative to the aggregate price index $\bar{\delta}/Y$ only, and firms need not consider the their individual competitors; changes in the aggregate price index are a sufficient statistic.

This inverse demand curve can be inverted into a per-capita demand curve for an individual variety. We denote the price elasticity of this demand curve for an individual variety by

$$\sigma(\frac{y}{Y}) = \frac{\Upsilon'(\frac{y}{Y})}{-\frac{y}{Y}\Upsilon''(\frac{y}{Y})}.$$

Firms. Each variety is supplied by a single firm seeking to maximize profits under monopolistic competition. Firms can enter to supply new varieties by incurring a fixed entry cost of f_e units of labor. Upon entry, a firms draw a type $\theta \in \mathbb{R}^+$ from a distribution with density $g(\theta)$ and cumulative distribution function $G(\theta)$. Each firm's productivity A_θ is an increasing function of its type θ. Having drawn its type, the firm then decides whether to produce or to exit. Production requires paying an overhead cost of f_o units of labor, with a constant marginal cost of $1/A_\theta$ units of labor per unit of the good sold. Finally, the firm decides what price to set, taking as given their residual demand curve.

The profit-maximizing price p_θ of a producing firm of type θ is a markup μ_θ over its marginal cost w/A_θ. Its per-capita quantity y_θ is the demand at that price. The price, markup, and per-capita quantity are determined implicitly by

$$p_\theta = \mu_\theta \frac{w}{A_\theta}, \quad y_\theta = y(p_\theta), \quad \text{and} \quad \mu_\theta = \mu(\frac{y_\theta}{Y}),$$

[2] Let $e(p_\omega, Y)$ be the expenditure function of a household as a function of the price of all varieties p_ω and utility/welfare Y (where the price of unavailable varieties is equal to ∞). Since preferences are homothetic, we can write $e(p_\omega, Y) = P^Y(p_\omega)Y$, where $P^Y(p_\omega)$ is the ideal price index.

where the markup function is given by the usual Lerner formula

$$\mu(\frac{y}{Y}) = \frac{1}{1 - \frac{1}{\sigma(\frac{y}{Y})}}.$$

A firm of type θ chooses to produce if, and only if,

$$Lp_\theta y_\theta \left(1 - \frac{1}{\mu_\theta}\right) \geq f_o w.$$

Hence, there is an endogenous cutoff θ^* such that firms of type $\theta \geq \theta^*$ decide to produce, and firms of type $\theta < \theta^*$ exit. The zero-profit condition, associated with entry, is then

$$\frac{1}{\Delta} \int_{\theta^*}^{\infty} \left[Lp_\theta y_\theta \left(1 - \frac{1}{\mu_\theta}\right) - f_o w \right] g(\theta) d\theta \geq f_e w.$$

The parameter Δ is introduced to allow the equations to represent a repeated version of the static model with an infinite number of periods $0, 1, \cdots, \infty$, where each producing firm has an exogenous probability Δ of being forced to exit in every period $t = 0, 1, \cdots, \infty$. In the absence of discounting, the expected net present value of profits is then given by the left hand side of the entry equation.

Equilibrium. The equilibrium concept is straightforward: consumers maximize utility taking prices as given; firms maximize profits taking prices other than their own and consumer welfare as given; and markets clear.

2.2 Summary of the Equilibrium Conditions

Since data is usually recorded in terms of sales rather than physical quantities, we restate the model's conditions in terms of sales. Define the sales share density

$$\lambda_\theta = \frac{Mp_\theta y_\theta}{w},$$

where M is the mass of entrants (intuitively, the number of copies of firms of type θ that enter in equilibrium). The sales share density is such that the aggregate sales share of entrants (as a fraction of income) with type in $(\theta, \theta + d\theta)$ is $\lambda_\theta g(\theta) d\theta$. Quantities per capita y_θ and relative prices p_θ / w can all be recovered from the sales share density λ_θ and markups μ_θ:

$$y_\theta = \frac{\lambda_\theta A_\theta}{\mu_\theta M} \quad \text{and} \quad \frac{p_\theta}{w} = \frac{\mu_\theta}{A_\theta}.$$

It follows that all the equilibrium conditions can also be written entirely in terms of the endogenous equilibrium variables M, Y, λ_θ, μ_θ, and exogenous parameters L, f, $f_e\Delta$, and A_θ.

Using the expectation conditional on survival (conditional on $\theta \geq \theta^*$), consumer welfare is

$$1 = M\mathbb{E}\left[\Upsilon(\frac{\lambda_\theta A_\theta}{\mu_\theta MY})\right],$$

restating the definition of consumer welfare per capita (1). Similarly, the free entry condition is

$$\frac{Mf_e\Delta}{(1-G(\theta^*))L} = \mathbb{E}\left[\lambda_\theta\left(1-\frac{1}{\mu_\theta}\right)-\frac{Mf_o}{L}\right],$$

restating that entry costs exactly offset the aggregate variable profit share net of the overhead costs. The selection condition is

$$\frac{Mf_o}{L} = \lambda_{\theta^*}\left(1-\frac{1}{\mu_{\theta^*}}\right),$$

restating that the overhead costs exactly offset the variable profit share for the marginal producer type θ^*. The individual markup equations is

$$\mu_\theta = \mu(\frac{\lambda_\theta A_\theta}{\mu_\theta MY}).$$

restating the Lerner condition. Individual demand is

$$\frac{\mu_\theta}{A_\theta} = \frac{\bar{\delta}}{Y}\Upsilon'(\frac{\lambda_\theta A_\theta}{\mu_\theta MY}).$$

Finally, the demand index equation is

$$\frac{1}{\bar{\delta}} = M\mathbb{E}\left[\frac{\lambda_\theta A_\theta}{\mu_\theta MY}\Upsilon'(\frac{\lambda_\theta A_\theta}{\mu_\theta MY})\right].$$

To streamline the exposition, we have made use of the following convention. For two variable $x_\theta > 0$ and z_θ, we define

$$\mathbb{E}_x[z_\theta] = \frac{\int_{\theta^*}^{\infty} x_\theta z_\theta \frac{g(\theta)}{1-G(\theta^*)}d\theta}{\int_{\theta^*}^{\infty} x_\theta \frac{g(\theta)}{1-G(\theta^*)}d\theta}.$$

We write $\mathbb{E}$ to denote $\mathbb{E}_x$ when $x_\theta = 1$ for all θ. The operator $\mathbb{E}_x$ operates a change of measure by putting more weight on types θ with higher values of x_θ. In the rest of the paper, we will often encounter $\mathbb{E}$, $\mathbb{E}_\lambda$, and $\mathbb{E}_{\lambda(1-1/\mu)}$, which respectively correspond to integrals with respect to the physical density, the sales share density, and the profit share density.

3 Central Concepts and Solution Strategy

In this section, we introduce some central concepts for the analysis to come and describe our solution strategy.

3.1 Markups, Pass-Throughs, and Infra-Marginal Surplus

In order to characterize changes in consumer welfare and in real GDP, we will need the following definitions for markups, pass-throughs, and infra-marginal surplus ratios.

Markups, pass-throughs, and Marshall's second law of demand. It is important to define a number of notions related to markups and their behavior. We have already defined the markup function $\mu(y/Y)$ and described its relation to the elasticity of the individual demand function

$$\mu(\frac{y}{Y}) = \frac{1}{1 - \frac{1}{\sigma(\frac{y}{Y})}}.$$

We define the individual (partial equilibrium) pass-through ρ_ω of a variety as

$$\rho_\omega = -d \log p_\omega / d \log A_\omega = 1 - d \log \mu_\omega / d \log A_\omega,$$

the elasticity of its price to its productivity, holding all other prices and consumer welfare fixed. This elasticity is necessarily positive, and can be computed by differentiating the individual demand equation $\mu(\frac{y}{Y})/A = \bar{\delta}\Upsilon'(\frac{y}{Y})/Y$ with respect to A, holding Y and $\bar{\delta}$ constant

$$\rho(\frac{y}{Y}) = \frac{1}{1 + \frac{\frac{y}{Y}\mu'(\frac{y}{Y})}{\mu(\frac{y}{Y})}\sigma(\frac{y}{Y})}.$$

The markup and pass-through of a variety of type θ are denoted by $\mu_\theta = \mu(y_\theta/Y)$ and $\rho_\theta = \rho(y_\theta/Y)$.

Marshall's weak second law of demand is the requirement that markups be increasing with productivity. It is well known that it is equivalent to the requirement that the individual demand curve be log concave.[3] Given that productivity is increasing

[3] See, for example, Melitz (2018). The illustration in Figure 1 satisfies Marshall's weak second law of

in the type of a firm, it is also equivalent to the requirement that μ_θ be increasing in θ, or equivalently,

$$\mu'(\frac{y}{Y}) \geq 0.$$

Marshall's strong second law of demand is the requirement that pass-throughs be decreasing with productivity. The strong law implies the weak law, and is equivalent to the requirement that individual marginal revenue curve be log concave. Given that productivity is increasing in the type of a firm, it is also equivalent the requirement that ρ_θ be decreasing in θ, or equivalently

$$\rho'(\frac{y}{Y}) \leq 0.$$

We do not impose Marshall's second law of demand. However, both the weak and strong version have some empirical support, for example Amiti et al. (2019), and turn out to be useful benchmarks for understanding the comparative statics of the model.

Infra-marginal surplus ratio and aligned preferences. We also define the infra-marginal surplus ratio of a variety δ as the amount of infra-marginal surplus per unit sales. More precisely, it is the ratio of the consumption equivalent utility $\bar{\delta}\Upsilon(y/Y)$ from a marginal variety to its sales share $py/w = \bar{\delta}(y/Y)\Upsilon'(y/Y)$. It is given by the infra-marginal surplus ratio function

$$\delta(\frac{y}{Y}) = \frac{\Upsilon(\frac{y}{Y})}{\frac{y}{Y}\Upsilon'(\frac{y}{Y})} \geq 1.$$

Mathematically, δ is the inverse of the returns to scale in Υ. The infra-marginal surplus ratio of a variety of type θ is denoted by $\delta_\theta = \delta(y_\theta/Y)$. Figure 2 depicts the graphical intuition for δ — it is the ratio of consumer surplus $A + B$ divided revenues by B.[4]

Note that the demand index is exactly the sales-weighted average of the infra-marginal surplus ratios[5]

$$\mathbb{E}_\lambda[\delta_\theta] = \bar{\delta}.$$

We say that (social and private) preferences are aligned if the infra-marginal consumer surplus ratio varies with productivity in the same direction as markups, or in

demand.

[4] When goods enter and exist at a choke price, as in Arkolakis et al. (2018), we naturally have $B = 0$ for entering firms. In these cases, the entry-exit margin can be said to be "neoclassical" in the sense that revenues reflect consumer surplus. As can be seen from footnote 6, in such models, the equivalence between real GDP and welfare is restored.

[5] This follows from the definition $\overline{\delta} = 1/\mathbb{E}(M\Upsilon'(y_\theta/Y)y_\theta/Y)$, and substituting $1 = \mathbb{E}(\Upsilon(y_\theta/Y))$.

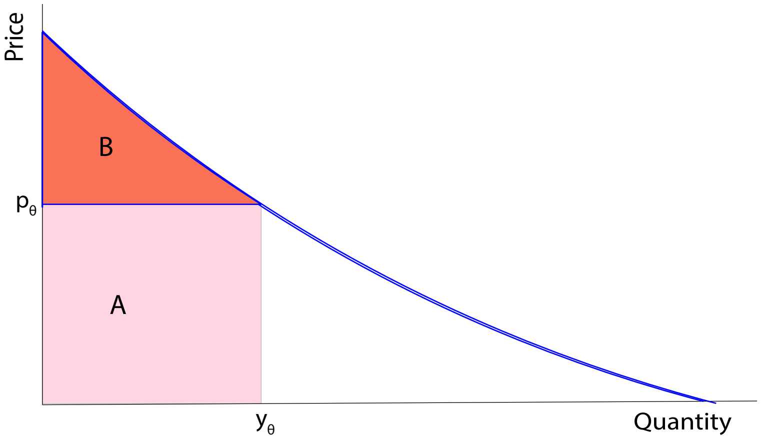


Figure 2: Graphical illustration of δ_θ.

other words if $\delta'(y/Y)$ and $\mu'(y/Y)$ have the same sign.[6] For example, if Marshall's weak second law of demand holds, so that markups increase with productivity and $\mu'(y/Y) \geq 0$, preferences are aligned if and only if δ also increases with productivity. It is easy to see that a necessary and sufficient condition in terms of demand primitives is

$$\delta'(\frac{y}{Y}) = \frac{1 - [\sigma(\frac{y}{Y}) - 1][\delta(\frac{y}{Y}) - 1]}{\frac{y}{Y}\sigma(\frac{y}{Y})} \geq 0. \tag{3}$$

3.2 Consumer Welfare and Real Output

We are interested in changes in consumer welfare and real output (real GDP) per capita in response to changes in the exogenous parameters. The change in consumer welfare is[7]

$$d \log Y.$$

[6] This terminology, due to Dhingra and Morrow (2019), captures the idea that when preferences are "aligned," then private gains (which are increasing in markups) are aligned with social preferences (which are increasing in the infra-marginal consumer surplus).

[7] This notion of consumer welfare coincides with equivalent variation. Let M_θ denote the mass of products of type θ, then in general we can write

$$d \log Y = -E_\lambda \left(d \log p_\theta / w \right) + E_\lambda \left((\delta_\theta - 1) d \log M_\theta \right).$$

The first term measures the marginal surplus from changes in prices, and absent changes in product variety ($d \log M_\theta$=0) is just Shephard's lemma. The second term measures the infra-marginal surplus from changes in variety. With knowledge of turnover in varieties $d \log M_\theta$ and the infra-marginal surplus ratios δ_θ, this equation could be used to measure changes in consumer welfare.

Changes in real GDP per capita are defined using idealized versions of the procedures that applied by statistical agencies. That is, using Divisia indices for continuing varieties present before and after the change. In principle, changes in real GDP can either be defined using the Divisia quantity index or the Divisia price index. In the body of the paper, we use the price index definition and include a discussion of the quantity index in Appendix E. Hence, the change in real GDP or output per capita is defined to be nominal income deflated by the GDP deflator. In other words,

$$d \log Q = -\mathbb{E}_{\lambda}[d \log(\frac{p_{\theta}}{w})].$$

Since the supply of the primary factor (labor) is exogenous in our model, changes in real GDP per capita $d \log Q$ are equal to changes in aggregate TFP (per capita). An important theme of this paper is that changes in welfare and changes in aggregate TFP are very different objects with very different determinants.

If this model did not allow for products creation and destruction, then changes in consumer welfare $d \log Y$, and changes in real output per capita $d \log Q$ would coincide.[8]

More concretely,

$$d \log Y = \left(\mathbb{E}_{\lambda}[\delta_{\theta}] - 1\right) d \log M + \left(\mathbb{E}_{\lambda}[\delta_{\theta}] - \delta_{\theta^*}\right) \lambda_{\theta^*} \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^* + \mathbb{E}_{\lambda}\left[d \log(\frac{A_{\theta}}{\mu_{\theta}})\right],$$

and

$$d \log Q = \mathbb{E}_{\lambda}\left[d \log(\frac{A_{\theta}}{\mu_{\theta}})\right].$$

Intuitively, consumer welfare changes $d \log Y$ incorporate the infra-marginal consumer surplus brought about by the entry of new varieties $d \log M$, or destroyed by the exit of varieties $d\theta^*$ via the first two terms on the right-hand side of the expression. By contrast, changes in real output per capita $d \log Q$ do not, and only take into account changes in the intensive margin of prices $d \log(p_{\theta}/w) = d \log(\mu_{\theta}/A_{\theta})$.

[8] This is also true in models of entry devoid of non-convexities featuring no fixed costs and demand curves with choke prices, where prices and quantities at the variety level change smoothly, for example Arkolakis et al. (2018). See Appendix E for more information on conditions under which real GDP per capita measures welfare.

3.3 Pure Changes in Technology and Changes in Allocative Efficiency

To understand changes in consumer welfare and real output in response to changes in exogenous parameters, it will be useful to decompose them into pure changes in technology and changes in allocative efficiency. Pure changes in technology capture the direct impact of the shock, holding the allocation of resources constant. Changes in allocative efficiency capture the indirect impact of the equilibrium reallocation of resources triggered by the shock. These are typically nonzero at the first order because the economy is not efficient to begin with.

To make this precise, following Baqaee and Farhi (2017b), we define the allocation vector $\mathcal{X} = (l_e, l_o, \{l_\theta\})$. It describes the fractions of labor allocated to the following activities: entry, overhead, and variable production of varieties of type θ. Together with the productivity vector $\mathcal{A} = (L, f_e\Delta, f_o, \{A_\theta\})$, it entirely describes any feasible allocation. Let $\mathcal{Y}(\mathcal{A}, \mathcal{X})$ be the associated level of consumer welfare.

We can decompose changes in consumer welfare into pure changes in technology and changes in allocative efficiency

$$d\log Y = \underbrace{\frac{\partial \log \mathcal{Y}}{\partial \log \mathcal{A}} d\log \mathcal{A}}_{\text{pure technology}} + \underbrace{\frac{\partial \log \mathcal{Y}}{\partial \mathcal{X}} d\mathcal{X}}_{\text{allocative efficiency}},$$

Pure changes in technology are the changes in consumer welfare or real output that are directly due to changes in technology $d\log\mathcal{A}$, holding the allocation of resources $\mathcal{X}$ constant. Changes in allocative efficiency are the changes in consumer welfare or real output that are due to the equilibrium reallocation of resources $d\mathcal{X} = (d\mathcal{X}/d\log\mathcal{A})d\log\mathcal{A}$ triggered by the shocks, holding technology $\mathcal{A}$ constant. In efficient economies, the envelope theorem implies there are only pure changes in technology and no changes in allocative efficiency. In inefficient economies, there are both pure changes in technology and changes in allocative efficiency.

3.4 Solution Strategy

In the following sections, we will provide analytical characterizations of first-order comparative statics of the model with respect to changes in the exogenous parameters. The representation of the equilibrium in Section 2.2 makes clear that such characterizations can be broken down into the following steps.

First, characterize changes in entry $d\log M$, markups $d\log\mu_\theta$, and selection cutoff $d\theta^*$, as a function of changes in consumer welfare $d\log Y$, using the free-entry condition, the selection condition, the individual markup equation, the individual demand equation, and the demand index equation. Second, aggregate these changes into changes in consumer welfare $d\log Y$ using the formulas in Section 3.2. Solve the resulting fixed point. Third aggregate these changes into changes in aggregate output per capita $d\log Q$ using the formula in Section 3.2. Fourth and finally, decompose these changes into pure changes in technology and changes in allocative efficiency along the lines of Section 3.3.

Non-parametric sufficient statistics. These characterizations will avoid putting any additional parametric structure on the model. For example, they will not impose any specific functional form on the Kimball aggregator or the productivity distribution. Instead, they will be expressed in terms of ex-ante measurable non-parametric sufficient statistics introduced in Section 3.1: sales shares λ_θ, markups μ_θ, pass-throughs ρ_θ, and relative infra-marginal consumer surplus ratios δ_θ. They will also depend on the hazard rate $\gamma^*_\theta = g_a(\log A_{\theta^*})/[1 - G_a(\log A_{\theta^*})]$ of the log-productivity distribution at the selection cutoff, where $g_a(\log A_\theta) = g(\theta)/(\partial \log A_\theta/\partial\theta)$. Will make contact with the data through these sufficient statistics.

4 Homogeneous Firms

To build intuition, we start by analyzing the case where firms are homogeneous. This case is obtained by assuming that all types have the same productivity $A_\theta = A$. We denote the common markup, pass-through, and individual demand elasticity by μ, ρ, and σ. For convenience, we normalize the wage to $w = 1$ throughout.

We proceed as follows. We start by discussing social inefficiency. We then study shocks to population, and end by analyzing the CES case. Shocks to other exogenous parameters, like productivity or fixed costs, are treated in Appendix C.

To aid with the intuition, we state our results using both markups μ and elasticities σ, but the two are of course connected via the Lerner condition $\mu = 1/(1 - 1/\sigma)$, or equivalently $\sigma = 1/(1 - 1/\mu)$.

4.1 Sources of Inefficiency

With homogeneous firms, the only margin that can be distorted is the allocation of labor to entry (and overhead) vs. variable production. As a result, social efficiency boils down to entry efficiency. This will no longer be true with heterogenous firms, where distortions can arise on several different margins.

The allocation matrix gives us an intuitive way to think about entry efficiency. Starting at the initial equilibrium, change the allocation of resources by increasing the fraction of labor allocated to entry and overhead and decreasing the fraction of labor allocated to variable production. Compute the resulting change in consumer welfare. We say that there is too much entry if the change in consumer welfare is negative and that there is too little entry if it is positive.[9] We will show that there is too much (too little) entry if and only if the following condition is verified (violated)

$$\delta < \mu.$$

Rearranging (3) shows this condition is automatically verified under weak second Marshall law of demand and aligned preferences.

To understand this result , we will apply the following formula, which can easily be obtained by simple differentiation of the consumer welfare definition

$$d \log Y = \delta d \log M + d \log y.$$

In turn, the intuition for this formula is straightforward: new varieties capture a sales share equal to $d \log M$, with an effect $\delta d \log M$ on consumer welfare; the per-capita quantity of each existing variety changes by $d \log y$, with an effect $d \log y$ on consumer welfare.

Note that the initial allocation of labor allocates a fraction $l = 1/\mu$ to variable production and $l_e + l_o = 1 - 1/\mu$ to entry and overhead. Consider a reduction in the fraction of labor allocated to variable production $d \log l < 0$ and a complementary increase in the fraction of labor allocated to entry and overhead $d \log l_e = d \log l_o = -[1/(\mu - 1)] d \log l > 0$. The change in consumer welfare is $d \log Y = [1 - (\delta - 1)/(\mu - 1)] d \log l$, which is negative (too much entry) if and only if $\delta < \mu$.

Indeed, in this experiment, we have $d \log y = d \log l - d \log M < 0$, since the amount of labor per capita available for each variety is reduced by the reallocation of labor away from variable production and by the labor required to produce the new varieties. We also have $d \log M = -1/(\mu - 1) d \log l > 0$, because of the reallocation of labor to entry and overhead. The result follows.

[9] Note that the comparative static underlying this definition is a feasible allocation, but not an equilibrium allocation. It can only be supported as an equilibrium allocation by introducing a subsidy on entry. The defining question can then be reformulated as whether this subsidy on entry decreases or increases equilibrium consumer welfare.

Another way to understand the result draws on the intuition in Mankiw and Whinston (1986). Whether or not there is too much or too little entry depends on the relative strength of two offsetting effects. First, there is the non-appropriability effect. It pushes in the direction of too little entry because entering firms do not internalize the infra-marginal consumer surplus that they create. Firm revenues do not reflect total consumer surplus. The non-appropriability effect is commensurate with the relative gap $\delta - 1$. It is stronger, the higher is the infra-marginal surplus ratio δ. Second, there is the business stealing effect. It pushes in the direction of too much entry. Entering firms steal revenues from incumbent firms. They do not internalize the corresponding loss of profits. The business stealing effect is commensurate with $\mu - 1$. There is too much entry if non-appropriability effect is weaker than the business stealing effect: $\delta - 1 < \mu - 1$. Conversely, there is too little entry if the non-appropriability effect is stronger than the business stealing effect: $\delta - 1 > \mu - 1$.

4.2 Shocks to Population

For brevity, we study shocks to population, and treat fixed costs and productivity shocks in Appendix C. Increases in population can be interpreted literally as immigration or increased fertility, or they can be interpreted, somewhat metaphorically, as the integration (through trade) of more and more countries which are otherwise operating under autarky. Changes in consumer welfare and real output in response to increases in population can therefore be interpreted as either gains from scale or as gains from trade.

Proposition 1. *Suppose that firms have the same productivity $A_\theta = A$. In response to changes in population $d \log L$, changes in consumer welfare are given by*

$$d \log Y = \underbrace{(\delta - 1) d \log L}_{\textit{pure technology}} + \underbrace{\delta \frac{\xi}{1 - \xi} d \log L}_{\textit{allocative efficiency}},$$

where

$$\xi = \left(1 - \rho\right)\left(1 - \frac{\delta - 1}{\mu - 1}\right)\frac{1}{\sigma} = \left(1 - \rho\right)\left(1 - \frac{\delta}{\mu}\right).$$

We assume throughout that $\xi < 1$, which guarantees that $0 < d \log Y < \infty$. The first expression for ξ is arguably more complex, but we list it here because it will be useful to understand the intuition and to compare with the case with heterogenous firms below.

In response to a positive population shock $d\log L > 0$, there are in general both pure changes in technology and changes in allocative efficiency. Pure changes in technology are given by $(\delta - 1)d\log L > 0$. Recall that the pure effect of technology holds fixed the fraction of labor allocated to entry, overhead, and variable production. Because we hold the fraction of labor allocated to entry constant, the increase in population implies a proportional increase $d\log L > 0$ in entry (and overhead). The sales share captured by these new varieties is also $d\log L$. Therefore the increase in the number of varieties increases consumer welfare by $\delta d\log L > 0$. On the other hand, the increase in the number of varieties reduces the amount of labor per capita allocated to the production of each variety, and hence the per-capita quantity of each variety, by $d\log L$. This implies a reduction $-d\log L < 0$ in consumer welfare. The overall effect balances out these two offsetting effects.

Changes in allocative efficiency are given by $\delta[\xi/(1-\xi)]d\log L$. Since $\xi < 1$, the shock increases consumer welfare $d\log Y > 0$. It reduces the price index by $d\log(\delta/Y) = -(1/\sigma)(d\log Y + d\log L) < 0$. This triggers a reduction in markups by $d\log\mu = (1-\rho)d\log(\delta/Y) < 0$. This reduces the variable profit share and hence entry by $[1/(\mu-1)](1-\rho)d\log(\delta/Y) < 0$. This in turn changes consumer welfare by $[(\delta-1)/(\mu-1)-1](1-\rho)d\log(\delta/Y)$. These changes in consumer welfare are positive if and only if there is too much entry to begin with ($\delta < \mu$). The result in the proposition is obtained by replacing $d\log(\delta/Y)$ for its expression as a function of $d\log Y$ and solving for the fixed point.

The fact that markups respond to market size is called the *pro-competitive* effect of market size. In the homogeneous firm case, these pro-competitive effects are the source of changes in allocative efficiency. In the next section, we see that in the presence of heterogeneity, there are other drivers of allocative efficiency that are not related to the pro-competitive effect. In fact, quantitatively, we find these much-talked about pro-competitive effects are quantitatively much less significant than the other drivers of reallocation.

Proposition 1 allows us to easily determine the sign of changes in allocative efficiency, and hence whether changes in allocative efficiency amplify or mitigate the effects of the shocks.

Corollary 1. *Suppose that firms have the same productivity $A_\theta = A$. Increased population increases allocative efficiency if and only if $\xi > 0$. As long as pass-through is incomplete ($\rho < 1$), this is equivalent to there being too much entry $\delta < \mu$. There is always too much entry under weak second Marshall law of demand and aligned preferences.*

Now, consider the effects of the population shock on real GDP per capita.

Proposition 2. *Suppose that firms have the same productivity $A_\theta = A$. In response to changes in population $d \log L$, changes in real output per capita are given by*

$$d \log Q^p = \frac{1-\rho}{\sigma}(d \log Y + d \log L),$$

where $d \log Y$ is given by Proposition 1.

Changes in real output per capita are $d \log Q = -d \log p$, an increase in population leads to a reduction in markups, and this this pro-competitive effect, in turn, increases increase real output.

Discussion. Under the assumptions of Corollary 1, an increase in population, which as discussed above can also be interpreted as an increase in trade integration, increases consumer welfare, through different channels.

The positive pure changes in technology arise from increasing economies of scale. When the allocation of resources is kept constant, the total quantity sold by each firm remains constant, the per-capita quantity declines, and new firms enter.

There are also positive changes in allocative efficiency which come into play as the allocation of resources adjusts to the shock. Because existing firms reduce their markups, the total quantities sold by these firms expands at the expense of entry. If there was too much entry to begin with (the infra-marginal surplus ratio is lower than the average markup), this reallocation of resources is beneficial.

The beneficial nature of these reallocation effects hinges entirely on second-best principles. It depends on whether there was too much or too little entry from a social perspective to begin with, and on whether reallocations decrease or increase entry.

An increase in population is always pro-competitive, in the sense, that it always reduces markups and expands the total (not per capita) quantities produced by existing firms. However, these pro-competitive effects do not necessarily increase consumer welfare. They do so only if there was too much entry to begin with, because the reduction in markups has the effect of reducing profit shares, and hence entry. By contrast, the effect on real GDP per capita is unambiguous, the reduction in markup always increases real GDP per capita because it is associated with decreases in prices of existing firms.

4.3 CES Example

It is interesting to study the CES benchmark, obtained by setting $\Upsilon(x) = x^{\frac{\sigma-1}{\sigma}}$, where with some abuse of notation, $\sigma > 1$ is some scalar. In this case, the elasticity of substitution is constant and equal to σ, markups are constant $\mu = 1/(1 - 1/\sigma)$, pass-throughs are equal to one $\rho = 1$, and the infra-marginal surplus ratio is constant $\delta = \sigma/(\sigma - 1)$. Moreover, entry is efficient since $\delta = \mu$.

Changes in consumer welfare are given by

$$d \log Y = \underbrace{(\delta - 1) d \log L}_{\text{pure technology}} + \underbrace{0}_{\text{allocative efficiency}} .$$

Unlike in the general case, shocks to population only lead to pure changes in technology and do not lead to any change in allocative efficiency, with an overall effect $d \log Y = (\delta - 1) d \log L$. The reason, which is twofold, is straightforward: first, there is no change in markups ($\rho = 1$) and hence no change in the fraction of resources dedicated to entry; second entry is actually efficient to begin with ($\delta = \mu$).

The response of real GDP per capita is

$$d \log Q = 0,$$

since markups and productivity shifters do not change. In Appendix C, we show that similar results hold for shocks to productivities and fixed costs (no changes in allocative efficiency).

5 Heterogenous Firms

In this section, we turn to the case of heterogenous firms. For convenience, we normalize the wage to $w = 1$ throughout. Before proceeding, we define the key notions of entry and selection efficiency.

We proceed as follows. We start by discussing social inefficiency. We then study shocks to population, and end by analyzing the CES case. Shocks to fixed costs and productivity are in Appendix C

To aid with the intuition, we state our results using both markups μ_θ and price-elasticities σ_θ, but the two are of course connected via the Lerner condition $\mu_\theta = 1/(1 - 1/\sigma_\theta)$, or equivalently $\sigma_\theta = 1/(1 - 1/\mu_\theta)$. Using this observation, our results can be expressed in terms of the following sufficient statistics: sales shares λ_θ, markups μ_θ, pass-throughs ρ_θ, the infra-marginal surplus ratio δ_θ, and the hazard rate of log productivity at the selection cutoff $\gamma_{\theta^*} = g_a(\log A_{\theta^*})/[1 - G_a(\log A_{\theta^*})]$, where $g_a(\log A_\theta) = g(\theta)/(\partial \log A_\theta / \partial \theta)$.

5.1 Sources of Inefficiency

With homogeneous firms, the only margin that could be distorted was the allocation of labor to entry (and overhead) vs. production. With heterogenous firms, more margins can be distorted: the allocation of labor to entry (and overhead) vs. production, but also the selection cutoff determining which varieties are allocated labor for variable production at all, and the allocation of labor for variable production across non-exiting varieties.

The allocation matrix continues to give us an intuitive way to think about efficiency along these different margins.[10] The following expression for changes in consumer welfare, derived from the definition of consumer welfare, will be useful for this purpose

$$d \log Y = \mathbb{E}_\lambda[\delta_\theta] d \log M - \delta_{\theta^*} \lambda_{\theta^*} \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^* + \mathbb{E}_\lambda[d \log y_\theta].$$

Entry efficiency. We say that there is too much (too little) entry if consumer welfare increases (decreases) when labor is reallocated from variable production to entry and overhead, but keeping the selection cutoff and the relative allocation of labor across non-exiting varieties constant. There is too much (too little) entry if and only if the following condition is verified (violated)[11]

$$\mathbb{E}_\lambda[\delta_\theta] < \mathbb{E}_\lambda \left[\mu_\theta^{-1}\right]^{-1},$$

which is a comparison of the sales-weighted average of the infra-marginal surplus ratios and the harmonic average of markups.

Note that the initial allocation of labor allocates a fraction $l = \mathbb{E}[l_\theta] = \mathbb{E}_\lambda[1/\mu_\theta]$ to variable production and $l_e + l_o = 1 - \mathbb{E}_\lambda[1/\mu]$ to entry and overhead. Consider a reduction in the fraction of labor allocated to variable production $d \log l_\theta = d \log l < 0$ and a complementary increase in the fraction of labor allocated to entry and overhead $d \log l_e = d \log l_o = -[\mathbb{E}_\lambda[1/\mu_\theta]/(1 - \mathbb{E}_\lambda[1/\mu_\theta])] d \log l > 0$. The reduction in the per-capita quantity of each variety $d \log y_\theta = d \log l - d \log M < 0$ reduces consumer welfare by $\mathbb{E}_\lambda[d \log y_\theta] < 0$. We also have an increase in entry $d \log M = -[\mathbb{E}_\lambda[1/\mu_\theta]/(1 - \mathbb{E}_\lambda[1/\mu_\theta])] d \log l > 0$, which increases consumer welfare by $\mathbb{E}_\lambda[\delta_\theta] d \log M > 0$. Finally,

[10] Once again, the comparative static underlying these definitions are feasible allocations, but not equilibrium allocations. They can only be supported as an equilibrium allocation by introducing taxes and subsidies. The defining question can then be reformulated as whether these taxes and subsidies decrease or increase equilibrium consumer welfare.

[11] Unlike in the case with homogeneous firms, this condition is no longer automatically verified under weak second Marshall law of demand and aligned preferences.

we have no change in selection $d\theta^* = 0$. The overall effect on consumer welfare is given by $d \log Y = [1 - (\mathbb{E}_\lambda[\delta_\theta] - 1)\mathbb{E}_\lambda[1/\mu_\theta]/(1 - \mathbb{E}_\lambda[1/\mu_\theta])] d \log l$, which is negative (too much entry) if and only if $1/\mathbb{E}_\lambda[1/\mu_\theta] < \mathbb{E}_\lambda[\delta_\theta]$. In the homogeneous firm case, this condition collapses to the simple $\delta < \mu$.

Selection efficiency. We say that there is too little (too much) selection if consumer welfare increases (decreases) when the selection cutoff is increased and the labor previously allocated to variable production and overhead of the newly exiting varieties is reallocated proportionately to entry, overhead, and to variable production. There is too little (too much) selection if and only if the following condition is verified (violated)

$$\delta_{\theta^*} < \mathbb{E}_\lambda[\delta_\theta].$$

Suppose that we increase the selection cutoff by $d\theta^* > 0$, and reallocate the labor previously allocated to the variable production and overhead of varieties with type in $[\theta^*, \theta^* + d\theta^*)$ proportionately to entry, overhead, and variable production. The exiting varieties reduce consumer welfare by $-\delta_{\theta^*}\lambda_{\theta^*}[g(\theta^*)/(1 - G(\theta^*))]d\theta^*$. The new varieties $d \log M = \lambda_{\theta^*}[g(\theta^*)/(1 - G(\theta^*))]d\theta^*$ increases consumer welfare by $\mathbb{E}_\lambda[\delta_\theta] d \log M$. There is no change in the production of existing varieties $d \log y_\theta = 0$. The overall effect on consumer welfare is $(\mathbb{E}_\lambda[\delta_\theta] - \delta_{\theta^*})\lambda_{\theta^*}[g(\theta^*)/(1 - G(\theta^*))]d\theta^*$, which is positive (too little selection) if and only if $\delta_{\theta^*} < \mathbb{E}_\lambda[\delta_\theta]$.

Relative production efficiency. Finally, we say that the supply of a variety is too large (too small) compared to another if consumer welfare increases (decreases) when labor is reallocated from the former to the latter. The supply of variety θ' is too large (too small) compared to that of variety θ if and only if the following condition is verified (violated)

$$\mu_{\theta'} < \mu_\theta.$$

Following Baqaee and Farhi (2017a), consider a reduction $d \log l_{\theta'} < 0$ in the fraction of labor allocated to the supply of varieties in $(\theta', \theta' + d\theta')$ and a complementary increase $d \log l_\theta = -(g(\theta')/g(\theta))(l_{\theta'}/l_\theta) d \log l_{\theta'} > 0$ in the fraction of labor allocated to the supply of varieties in $(\theta, \theta + d\theta')$, which, using the fact that $l_{\theta'}/l_\theta = (\lambda_{\theta'}/\mu_{\theta'})/(\lambda_\theta/\mu_\theta)$, can be rewritten as $d \log l_\theta = -(g(\theta')/g(\theta))(\lambda_{\theta'}/\mu_{\theta'})/(\lambda_\theta/\mu_\theta) d \log l_{\theta'} > 0$. This leads to an decrease $d \log y_{\theta'} = d \log l_{\theta'} < 0$ in the quantity of the former varieties and an increase $d \log y_\theta = -(g(\theta')/g(\theta))(\lambda_{\theta'}/\mu_{\theta'})/(\lambda_\theta/\mu_\theta) d \log l_{\theta'} > 0$ in the quantity of the latter varieties. This effect on consumer welfare is $g(\theta')\lambda_{\theta'} d \log y_{\theta'} d\theta' + g(\theta)\lambda_\theta d \log y_\theta d\theta' = -(\mu_\theta/\mu_{\theta'} - 1)\lambda_{\theta'} g(\theta') d\theta' d \log l_{\theta'}$, which is positive if and only $\mu_\theta > \mu_{\theta'}$.

5.2 Shocks to Population

Consider shocks to population. As before, increases in population can either be interpreted as gains from scale, or under some assumptions, as gains from trade.

Proposition 3. *In response to changes in population $d\log L$, changes in consumer welfare are given by*

$$d\log Y = \underbrace{\left(\mathbb{E}_\lambda[\delta_\theta] - 1\right) d\log L}_{\text{pure technology}} + \underbrace{\frac{\xi^\epsilon + \xi^\mu + \xi^{\theta^*}}{1 - \xi^\epsilon - \xi^\mu - \xi^{\theta^*}} \left(\mathbb{E}_\lambda[\delta_\theta]\right) d\log L}_{\text{allocative efficiency}},$$

where

$$\begin{aligned}
\xi^\epsilon &= \left(\mathbb{E}_\lambda[\delta_\theta] - 1\right)\left(\mathbb{E}_\lambda[\sigma_\theta] - \mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta]\right)\left(\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\right),\\
\xi^{\theta^*} &= \left(\mathbb{E}_\lambda[\delta_\theta] - \delta_{\theta^*}\right)\left(\lambda_{\theta^*}\gamma_{\theta^*}\frac{\sigma_{\theta^*} - \mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta]}{\sigma_{\theta^*} - 1}\right)\left(\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\right),\\
\xi^\mu &= \left(\mathbb{E}_\lambda\left[(1-\rho_\theta)\left(1 - \frac{\mathbb{E}_\lambda[\delta_\theta] - 1}{\mu_\theta - 1}\right)\right]\right)\left(\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\right).
\end{aligned}$$

We assume throughout that $\xi^\epsilon + \xi^\mu + \xi^{\theta^*} < 1$, which guarantees that $0 < d\log Y < \infty$. We examine this expression term-by-term and explain its intuition.

The intuition for the pure changes in technology is the same as in the case of homogeneous firms covered in Section 4. The only difference is that the infra-marginal surplus ratios δ_θ are heterogenous and matter through their average $\mathbb{E}_\lambda[\delta_\theta]$. We now discuss the different changes in allocative efficiency associated with these different equilibria.

Each of ξ^ϵ, ξ^μ, and ξ^{θ^*} relate to adjustments along a specific margin. Consider an increase in population $d\log L > 0$, starting at the initial equilibrium. Each firm can adjust on three margins: its entry behavior; its exit behavior; and its price/markup. We decompose the general equilibrium response by analyzing three successive equilibrium allocations which allow firms to adjust along more and more margins. All three equilibrium allocations feature the same pure technology effect, but different changes in allocative efficiency, driven by different changes in the allocation of resources.[12]

Entry Only margin. First, consider the equilibrium where firms can only adjust their entry behavior (free entry) but can neither adjust their markups nor their exit

[12] In contrast to the non-equilibrium feasible counterfactuals underlying the definition of entry, selection, and production efficiency that we outlined in Section 5.1, these are equilibrium allocations.

behavior. Call welfare under this allocation Y^{ε}. The resulting changes in consumer welfare are given by Proposition 3 but with $\xi^{\mu} = \xi^{\theta^*} = 0$. They are strictly positive ($\xi^{\epsilon} > 0$) as long as there is non trivial heterogeneity, which we assume. The reduction in the price index triggers bigger reductions in per-capita quantities and sales for firms with higher price-elasticities and lower markups, which were too large to begin with compared to the firms with lower elasticities and higher markups. The associated reallocation towards high markup firms increases the variable profit share and entry. This is the effect that is graphically illustrated in Figure 1 in the introduction.

The detailed intuition for the changes in allocative efficiency captured by ξ^{ϵ} is as follows. The shock increases consumer welfare $d\log Y > 0$. It leads to a reduction in the price index by $d\log(\bar{\delta}/Y) = -\mathbb{E}_{\lambda}[1/\sigma_{\theta}](d\log Y + d\log L) < 0$. As a result, sales shares change by $(\sigma_{\theta} - \mathbb{E}_{\lambda}[\sigma_{\theta}])d\log(\bar{\delta}/Y)$. This reallocates resources towards varieties with lower elasticities σ_{θ}, which also have higher markups μ_{θ}, and increases the aggregate variable profit share and entry by $-(\mathbb{E}_{\lambda}[\sigma_{\theta}] - \mathbb{E}_{\lambda(1-1/\mu)}[\sigma_{\theta}])d\log(\bar{\delta}/Y) > 0$. This then increases consumer welfare by $-(\mathbb{E}_{\lambda}[\delta_{\theta}] - 1)(\mathbb{E}_{\lambda}[\sigma_{\theta}] - \mathbb{E}_{\lambda(1-1/\mu)}[\sigma_{\theta}])d\log(\bar{\delta}/Y) > 0$. The result in the proposition is obtained by replacing $d\log(\bar{\delta}/Y)$ by its expression as a function of $d\log Y$ and solving for the fixed point.

Entry and selection only margins. Second, consider the equilibrium where firms can adjust their entry behavior, but also their exit behavior (that is, firms can choose whether or not to exit after drawing their type). However, firms' markups by type stay constant. Denote welfare under this allocation by Y^{ε,θ^*}. The resulting changes in consumer welfare are given by Proposition 3 but with $\xi^{\mu} = 0$. There is a new source of changes in allocative efficiency captured by $\xi^{\theta^*} \neq 0$.

Suppose that the weak second Marshall law of demand holds (markups are increasing in productivity). As we will see, this guarantees that the selection cutoff increases $d\theta^* > 0$. The sales shares of the newly exiting varieties with $\theta \in [\theta^*, \theta^* + d\theta^*)$ is $\lambda_{\theta^*}(g(\theta^*)/[1 - G(\theta^*)])d\theta^*$. It is equal to $\lambda_{\theta^*}\gamma_{\theta^*}d\log A_{\theta^*}$, where $d\log A_{\theta^*}$ is the change in productivity associated with a change in type from θ^* to $\theta^* + d\theta^*$. This reallocates sales from exiting firms to the average surviving firm and changes consumer welfare by $(\mathbb{E}[\delta_{\theta}] - \delta_{\theta^*})\lambda_{\theta^*}(g(\theta^*)/[1 - G(\theta^*)])d\theta^*$. These changes in allocative efficiency are positive ($\xi^{\theta^*} > 0$) if there is too little selection to begin with ($\mathbb{E}[\delta_{\theta}] > \delta_{\theta^*}$). By definition, this is guaranteed if the weak second Marshall law of demand holds and preferences are aligned.

It is important to note that the fact that θ^* increases is not, on its own, evidence of an improvement in allocative efficiency. In other words, increases in the cutoff θ^*, due to intensifying competition, are *only* socially desirable if the marginal firm provides households with less infra-marginal surplus than the average surviving firm. In fact, in our empirical Section 7, we find evidence against this idea, since we find that increases in the cutoff θ^* reduce welfare.

To complete the intuition, we now discuss the change in the selection cutoff $d\theta^*$. It can be shown that the change in variable profits at the selection cutoff is given by $(\sigma_{\theta^*} - \mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta])d\log(\bar{\delta}/Y)$, where the change in the price index is given by $d\log(\bar{\delta}/Y) = -\mathbb{E}_\lambda[1/\sigma_\theta](d\log Y + d\log L) < 0$.[13] Variable profits at θ^* decrease as long as the marginal firm is more price elastic than the average firm $\sigma_{\theta^*} > \mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta]$, which is guaranteed under Marshall's weak second law of demand. Under this assumption, the reduction in variable profits at the cutoff requires an offsetting increase in the selection cutoff $d\theta^* > 0$ so that productivity increases by $d\log A_{\theta^*} = -[1/(\sigma_{\theta^*}-1)](\sigma_{\theta^*} - \mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta])d\log(\bar{\delta}/Y) > 0$. The result in the proposition is obtained by replacing $d\log(\bar{\delta}/Y)$ by its expression as a function of $d\log Y$ and solving for the fixed point.

Entry, exit, and pricing/markup margins. Last, consider the equilibrium where firms can not only adjust their entry and exit behavior, but also their pricing/markup behavior — the change in welfare under this allocation $d\log Y$ is the full-blown decentralized equilibrium outcome. The resulting changes in consumer welfare are given by Proposition 3. There is a new source of changes in allocative efficiency captured by $\xi^\mu \neq 0$. Adjustments along this margin (markups) are the source of pro-competitive effects from market size.

The intuition for the additional changes in allocative efficiency captured by ξ^μ is very similar to that presented in the case of homogeneous firms in Section 4. The only difference is that the terms in ξ^μ are appropriately averaged versions of the now heterogenous underlying sufficient statistics. In the homogeneous firm case, the ξ^ϵ and ξ^{θ^*} were equal to zero, so we only had to contend with the markup margin. In that case, the reduction in markups (caused by increased entry) increased welfare if and only if there was too much entry to begin. In the heterogeneous firm case, these changes in allocative efficiency are positive ($\xi^\mu > 0$) if and only if $\mathbb{E}_{\lambda(1-\rho)}[1 - (\mathbb{E}_\lambda[\delta_\theta] - 1)/(\mu_\theta - 1)] > 0$. It is not clear in general if this condition is weaker or stronger than the condition that there is too much entry.

The reason is subtle. There is a general reduction in markups, which reduces entry, and increases consumer welfare if and only if there is too much entry. But there is also a bigger reduction in markups for firms with low pass-throughs, which under strong second Marshall law of demand, also have lower elasticities and higher markups. That

[13] It is also proportional to a positive term which increases with $\mathbb{E}_\lambda[\delta_\theta]$.

they have lower pass-throughs pushes for a reallocation of resources towards them, but that they have lower price-elasticities pushes in the other direction. Whether or not resources are reallocated towards these high-markup firms, and hence whether or not the associated reallocation effects increase or decrease consumer welfare, depends on whether or not the pass-through effect dominates the elasticity effect. In the former case excessive entry is a sufficient condition for $\xi^{\mu} > 0$ but not in the latter case.

When all three margins $\xi^{\epsilon}, \xi^{\theta^*}$, and $\xi^{\mu} \geq 0$, we can sign the change in allocative efficiency in response to a population shock.

Corollary 2. *Sufficient conditions for positive changes in allocative efficiency in response to increases in population are: (1) that the weak and strong second Marshall laws of demand hold; (2) that $\mathbb{E}_{\lambda}\left[(1-\rho_{\theta})\left[1-(\mathbb{E}_{\lambda}[\delta_{\theta}]-1)/(\mu_{\theta}-1)\right]\right] > 0$ which can be stronger or weaker than the condition for excessive entry; and (3) that there be too little selection $\mathbb{E}_{\lambda}[\delta_{\theta}] > \delta_{\theta^*}$, which is automatically verified if (1) holds and if, in addition, preferences are aligned. More precisely, we always have $\xi^{\epsilon} \geq 0$, with a strict inequality as long as there is non trivial heterogeneity; (1) and (3) imply $\xi^{\theta^*} > 0$; and (2) implies that $\xi^{\mu} > 0$.*

Finally, we can also characterize the change in real GDP per capita, which depends only on how the prices of existing goods change.

Proposition 4. *In response to changes in population $d\log L$, changes in real output per capita are*

$$d\log Q^{p} = \Big(\mathbb{E}_{\lambda}\left[(1-\rho_{\theta})\right]\Big)\Big(\mathbb{E}_{\lambda}\left[\frac{1}{\sigma_{\theta}}\right]\Big)\Big(d\log Y + d\log L\Big),$$

where $d\log Y$ is given by Proposition 3.

The result is basically an appropriately averaged version of that presented in the case of homogeneous firms in Section 4.

5.3 CES Example

Once again, we consider the CES benchmark, $\Upsilon(x) = x^{\frac{\sigma-1}{\sigma}}$, where $\sigma > 1$ is some scalar. With heterogeneous firms, this example is a closed-economy version of Melitz (2003). The markups are constant $\mu = 1/(1-1/\sigma)$, pass-throughs are equal to one $\rho = 1$, and the infra-marginal surplus ratio is constant $\delta = \sigma/(\sigma-1)$. Moreover, entry is efficient since $\bar{\delta} = \bar{\mu}$. Furthermore, since $\mathbb{E}_{\lambda}(\delta_{\theta}) = \delta_{\theta^*}$, the exit/selection margin is also efficient. Finally, since markups are constant, there is no adjustment of markups. This means that changes in consumer welfare are simply given by pure technology effects

$$d\log Y = \underbrace{(\delta - 1)\, d\log L}_{\text{pure technology}} + \underbrace{0}_{\text{allocative efficiency}} .$$

The fact that changes in allocative efficiency are zero is not surprising; the CES model is efficient, so this result is a consequence of the envelope theorem. Next, consider changes in real GDP per capita, which are

$$d\log Q = 0,$$

since markups and productivity shifters do not change in response to changes in population, despite the fact that welfare increases. In Appendix C, we show that in the CES benchmark, real GDP is constant in response to shocks to fixed and overhead costs as well. This follows from the fact that in the CES world, these shocks do not change markups or productivity shifters.

Discussion. We can now take stock and dispel some deeply ingrained misconceptions. Although we conduct this discussion for shocks to population, the spirit of our remarks applies more broadly to shocks to fixed costs and to productivities.[14] As we have already discussed above, an increase in population can also be interpreted as an increase in trade integration. Under the assumptions of Corollary 2, it increases consumer welfare, through different channels.

There are positive pure changes in technology arising from economies of scale because fixed entry and overhead costs can be spread across a larger population. When the allocation of resources is kept constant, the total quantity sold by each firm remains constant, the per-capita quantity declines, and new firms enter.

There are also positive changes in allocative efficiency which come into play as the allocation of resources adjusts to the shock. First, holding exit and pricing/markup behavior constant, the total quantity sold by large firms expands and that of small firms shrinks (because demand for their goods is more elastic, and drops more rapidly in response to increased entry).[15] Since large firms are too small to begin with compared small firms (the former charge higher markups than the latter), this reallocation of resources is beneficial.

[14] See Appendix C for formal results about shocks to fixed costs and productivities

[15] To be precise, holding fixed markups and the selection margin, the change in the total quantity of type-θ goods is $d\log y_\theta M = (\sigma_\theta - E_\lambda \sigma_\theta) d\log \frac{\bar{\delta}}{Y}$, where $d\log \frac{\bar{\delta}}{Y} < 0$ is the price index for substitution.

Second, holding pricing/markup behavior constant, the smallest firms exit and make room for new firms. If there was too little selection to begin with (exiting firms had lower infra-marginal consumer surplus ratios than average), this reallocation of resources is beneficial.

Third, because existing firms reduce their markups, the total quantities sold by these firms expands at the expense of entry. This last effect is the only that operates in models with homogeneous firms. If there was too much entry to begin with (the average infra-marginal surplus ratio is lower than the average markup), this reallocation of resources is beneficial. However, this beneficial expansion of existing firms at the expense of entry is complicated by ambiguous reallocation effects across existing firms: on the one hand, large firms reduce their markups more than small firms which tends to reallocate resources towards larger firms; on the other hand, large firms are less elastic so that a given markup reduction induces less reallocation towards them than for small firms. The former effect is beneficial but the latter is detrimental.

It is important to stress that the beneficial nature of these three reallocation effects hinges entirely on second-best principles: in which direction the underlying margin is distorted, and in which direction the reallocation effect is moving this margin. It has nothing to do with the productivities of expanding and shrinking or disappearing firms per se. Such misleading intuitions are routinely invoked in economic writings, for example when discussing the popular model of Melitz (2003). In particular, that model has CES preferences and is therefore efficient. As a result, reallocations of resources and movements in the selection cut off have no impact on consumer welfare (or GDP and TFP) to the first order.

The expansion of large firms at the expense of small firms, underlying the reallocation effects ξ^{ϵ} increases consumer welfare only because large firms were too small to begin with from a social perspective since they charge higher markups, not because they have "higher productivities".[16] Similarly, the exit of the smallest firms underlying the second reallocation effect ξ^{θ^*} increases consumer welfare only if selection was too weak to begin with, which amounts to assuming that the marginal firm has lower infra-marginal consumer surplus ratio than average. In fact, in our empirical application, we find that this condition is violated, so that increases in the productivity cutoff reduce welfare. Finally, the expansion of existing firms at the expense of entry underlying part of the third reallocation effect increases consumer welfare if existing firms are too small to begin with, which is the case when the average infra-marginal

[16] In fact, since firms are producing differentiated varieties, it makes little sense to try to compare the "level" of their productivity; their output is not measured in comparable units.

consumer surplus ratio is less than the average markup.

Furthermore, these reallocations affect consumer welfare and aggregate TFP through very different channels. For example, holding pricing/markups constant, measured aggregate TFP is completely unaffected. In other words, the nature of the consumer welfare gains is distinct from that of aggregate TFP gains. Again, the two notions are routinely conflated in economic writings, for example in discussions of the Melitz model.

6 Distance to Efficient Frontier

At this point, we can calculate the social costs of the distortions by considering the optimal allocation and approximating the losses around this allocation. To do this, imagine a social planner who can implement the efficient allocation by regulating markups and imposing sales taxes. A sufficient condition is to set markups according to the infra-marginal surplus each firm generates $\mu_\theta^{opt} = \delta_\theta$ and sales taxes to be the reciprocal of markups $\tau_\theta^{opt} = 1/\mu_\theta$. The markups provide socially optimal incentives along the extensive margin and the output taxes undo the inefficiencies brought about by dispersed markups. See Edmond et al. (2018) for an alternative implementation of the optimal allocation using taxes.[17] This section contributes to the literature by providing an analytical approximation for distance to the efficient frontier.

At the decentralized monopolistically competitive equilibrium, we instead have $\mu_\theta = (1 - 1/\sigma_\theta)^{-1}$ and $\tau_\theta = 1$. The proposition below provides a second-order approximation of the distance to the efficient frontier, providing a link between our framework and the literature on the social costs of misallocation (in particular, to Epifani and Gancia, 2011).

Proposition 5. *The difference between welfare at the first-best allocation and the decentralized equilibrium is approximately*

$$\log Y^{opt} - \log Y \approx \frac{1}{2}\mathbb{E}_\lambda\left[\sigma_\theta\left(\frac{\mu_\theta}{\mathbb{E}_\lambda\left[\delta_\theta\right]} - 1\right)^2\right] + \frac{1}{2}\lambda_{\theta^*}\gamma_{\theta^*}\left(\mathbb{E}_\lambda\left[\delta_\theta\right] - \delta_{\theta^*}\right)^2,$$

where the remainder term is order $\log(\mu/\mu^{opt})^3$ *and* $\log(\tau/\tau^{opt})^3$.

The first summand, capturing distortions amongst surviving firms, scales with the price elasticity σ_θ and the dispersion of markups μ_θ relative to the average infra-

[17] Bilbiie et al. (2019) also consider related issues in a dynamic context.

marginal consumption surplus $E_\lambda(\delta_\theta)$. In the CES case, σ_θ, μ_θ, and δ_θ are all constant and $\mu_\theta = \delta_\theta$, which means that losses are zero.

The second summand captures the distortions along the selection margin, and scales in the difference between the infra-marginal surplus of the marginal firm and that of the average. It also scales with the hazard rate of the log productivity distribution for the marginal firm γ_θ^*. If there are many firms at the cutoff (high λ_{θ^*}) or the cutoff moves very quickly (high γ_{θ^*}) in response to distortions, then the losses from selection inefficiency $\delta_{\theta^*} \neq \mathbb{E}_\lambda(\delta_\theta)$ are amplified.

In the homogeneous firm case,

$$\log Y^{opt} - \log Y \approx \frac{1}{2}\sigma\mathbb{E}\left[\left(\frac{\mu}{\delta} - 1\right)^2\right],$$

which simply depends on the gap between the markup and the infra-marginal consumption surplus ratio.

7 Empirical Application

In this section, we take the theory to the data. We first present our non-parametric model estimation procedure. We then implement it using Belgian data and present some counterfactuals.

7.1 Non-Parametric Model Estimation Approach

We start by describing our non-parametric estimation procedure. The key step of our approach is a procedure to derive a non-parametric estimate of the Kimball aggregator Υ. The construction will use the restrictions imposed by Kimball demand. Essentially, a bigger firm is a smaller firm that received a positive productivity shock. The time-series pass-through, which encodes how the markup of a firm responds to a productivity shock, is equal to the cross-sectional pass-through, which encodes how markups increase as we move up the productivity distribution.

Given two key pieces of information, we find the Kimball aggregator Υ that rationalizes the data. We take as given: (1) the density of sales shares λ_θ, and (2) the pass-through function ρ_θ. Since pass-throughs are third derivatives of the Kimball aggregator, we can recover Υ by solving a series of differential equations. For boundary conditions, we need to take a stand on the average levels of first and second derivatives, i.e. on the average markup $\bar{\mu}$ and on the average infra-marginal consumption surplus ratio $\bar{\delta}$ (these will be constants of integration). We will present our estimates for different values of these variables.

Observation of sales λ_θ and pass-throughs ρ_θ will allow us to back out productivities A_θ up to a normalizing constant and markups μ_θ given the average markup $\bar{\mu}$. Using $\sigma_\theta = 1/(1 - 1/\mu_\theta)$ to recover elasticities will then allow us to back out infra-marginal surplus ratios δ_θ and the whole Kimball aggregator up to the average infra-marginal surplus ratio $\bar{\delta}$. Basically, cross-sectional observations on pass-throughs allows us to trace the individual demand curve and hence to back out the Kimball aggregator Υ up to some constants $\bar{\delta}$ and $\bar{\mu}$. Through this procedure, we will therefore be able to recover the whole nonlinear structure of the model.[18]

In our empirical application, we will use a uniform type distribution with $g(\theta) = 1$ and $G(\theta) = 1 - \theta$ by ranking firms by increasing size and associating their type to the fraction of firms with lower sales. The type distribution itself is irrelevant: the only thing that matters is the relation of the measure over types to the measure over sales.

In principle, one could use markups μ_θ or infra-marginal surplus ratios δ_θ to recover Υ. However, these objects are much harder to estimate, requiring either strong structural assumptions in the case of μ_θ, and in the case of δ_θ experimental data tracing out individual demand curves. In comparative terms, estimating the pass-through function is less daunting (we use estimates from Amiti et al., 2019), but the downside is that it will require some outside information to pin down $\bar{\mu}$ and $\bar{\delta}$.

Productivities, quantities, elasticities, and infra-marginal consumption surplus ratios. Productivities A_θ and markups μ_θ must simultaneously solve the two differential equations

$$\frac{d\log\lambda_\theta}{d\theta} = \frac{\rho_\theta}{\mu_\theta - 1}\frac{d\log A_\theta}{d\theta},$$

$$\frac{d\log\mu_\theta}{d\theta} = (1 - \rho_\theta)\frac{d\log A_\theta}{d\theta}.$$

The intuition for the first differential equation for sales shares is the following: compared to a firm of type θ, a firm with type $\theta + d\theta$ has higher productivity $\log A_{\theta+d\theta} - \log A_\theta = d\log A_\theta/d\theta$, lower price $\log p_{\theta+d\theta} - \log p_\theta = \rho_\theta d\log A_\theta/d\theta$, and higher sales $\log\lambda_{\theta+d\theta} - \log\lambda_\theta = (\sigma_\theta - 1)\rho_\theta d\log A_\theta/d\theta$ with $\sigma_\theta - 1 = 1/(\mu_\theta - 1)$. The intuition for

[18] We refer the reader to Appendix D for the discussion of a model with taste shocks and heterogeneous overhead costs in which cross-section and times-series are entirely disconnected. Non-parametric estimation of this richer model requires additional data on markups μ_θ as well as taking a stand on the distribution of infra-marginal surplus ratios δ_θ and overhead costs $f_{o,\theta}$. And even then, only local non-parametric estimation can be achieved, which is just enough to allow the computation of local counterfactuals along the lines of the formulas presented in the paper.

the second differential equation for markups is as follows: compared to a firm of type θ, a firm with type $\theta + d\theta$ has higher markup $\log \mu_{\theta+d\theta} - \log \mu_\theta = (1-\rho_\theta) d \log A_\theta / d\theta$.

Combining the two differential equations yields

$$\frac{d \log \mu_\theta}{d\theta} = (\mu_\theta - 1) \frac{1-\rho_\theta}{\rho_\theta} \frac{d \log \lambda_\theta}{d\theta}.$$

Given sales shares λ_θ and pass-throughs ρ_θ, this differential equation allows us to recover markups μ_θ up to a constant μ_{θ^*}. The constant $\mu_{\theta^*} \geq 1$ can be chosen to match a given value of the (harmonic) sales-weighted average markup $\bar{\mu} \geq 1$.

Either of the two differential equations for sales shares or markups then allows us to recover productivities up to a constant A_{θ^*}. This constant A_{θ^*} can be normalized to 1. For example, using the differential equation for sales shares, we get

$$\frac{d \log A_\theta}{d\theta} = \frac{\mu_\theta - 1}{\rho_\theta} \frac{d \log \lambda_\theta}{d\theta},$$

with initial condition $A_{\theta^*} = 1$.

Next we can then recover quantities using

$$y_\theta = \frac{\lambda_\theta A_\theta}{M \mu_\theta}.$$

Quantities are increasing in θ since $d \log y_\theta / d \log \theta = \mu_\theta d \log \lambda_\theta / d \log \theta$. We denote by $\theta(y)$ the reverse mapping giving a firm type as a function of the quantity that it produces.

Finally, we can recover infra-marginal consumption surplus ratios using the differential equation

$$\frac{d \log \delta_\theta}{d\theta} = \frac{\mu_\theta - \delta_\theta}{\delta_\theta} \frac{d \log \lambda_\theta}{d\theta},$$

with initial condition δ_{θ^*} chosen such that $\mathbb{E}_\lambda[\delta_\theta] = \bar{\delta}$.[19]

Kimball aggregator. We can then recover the Kimball aggregator by combining the definition of δ_θ with the residual demand curve

$$\Upsilon(\frac{y}{Y}) = \frac{\lambda_{\theta(y)} \delta_{\theta(y)}}{\bar{\delta}}.$$

[19] It turns out that these differential equations can actually be solved in closed-form. The closed-form expressions are provided in Appendix F.

Fixed costs and Value of Cutoff. The information so far does not reveal the cutoff value θ^*, so calibrating this number requires outside information. To calibrate the marginal type θ^*, we step slightly outside the model and imagine that new firms operate for one year before they choose to shut down. Hence, in their first year, the unconditional probability of exit is higher than the exogenous death rate. We then fit a quasi-hyperbolic process to estimates of firm exit probability by age as reported by Pugsley et al. (2018). Condition on θ^*, we can back out the fixed costs using the free-entry condition

$$\frac{f_e \Delta}{L} + (1 - G(\theta^*))\frac{f_o}{L} = \frac{1}{M}\mathbb{E}\left[\lambda_\theta\left(1 - \frac{1}{\mu_\theta}\right)\right],$$

and the selection condition

$$\frac{f_o}{L} = \frac{1}{M}\lambda_{\theta^*}\left(1 - \frac{1}{\mu_{\theta^*}}\right),$$

where total population L can be normalized to 1, and Δ.

7.2 Empirical Implementation

In this section, we implement the formulas above using estimates of the firm-level pass-through function and distribution of sales. We consider shocks to population and fixed costs, and find that increases in population trigger large positive changes in allocative efficiency. Contrary to what one might expect, these are due to the fact that entry of new firms reallocates resources amongst firm types, and not due to the fact that entry reduces markups or that entry increases the selection cutoff. In fact, the fact that entry triggers changes in markups and increases the type of the marginal entrant counteract the positive effect.

Data sources. Here, we give a brief account of our data sources and procedures, and the full details can be found in Appendix A. We rely on Amiti et al. (2019) who report estimates of pass-throughs by firm size for manufacturing firms in Belgium. They use annual administrative firm-product level data (Prodcom) from 1995-2007, which contains information on prices and sales, collected by Statistics Belgium. They merge this with Customs data, and using exchange rate shocks as instruments for changes in marginal cost, they show that they can identify the partial equilibrium pass-through by firm size (under assumptions that are consistent with our model).

Prodcom does not sample very small firms (firms must have sales greater than 1 million euros to be included). Therefore, we merge their estimates of the pass-through function ρ (as a function of size) with the sales distribution λ for the universe of Belgian manufacturing firms (from VAT declarations). For firms that are smaller than the smallest firms in Prodcom, we assume that their pass-through is equal to one. This is consistent with the estimates of Amiti et al. (2019) who find that the average pass-through for the smallest 75% of firms in Prodcom is 0.97.

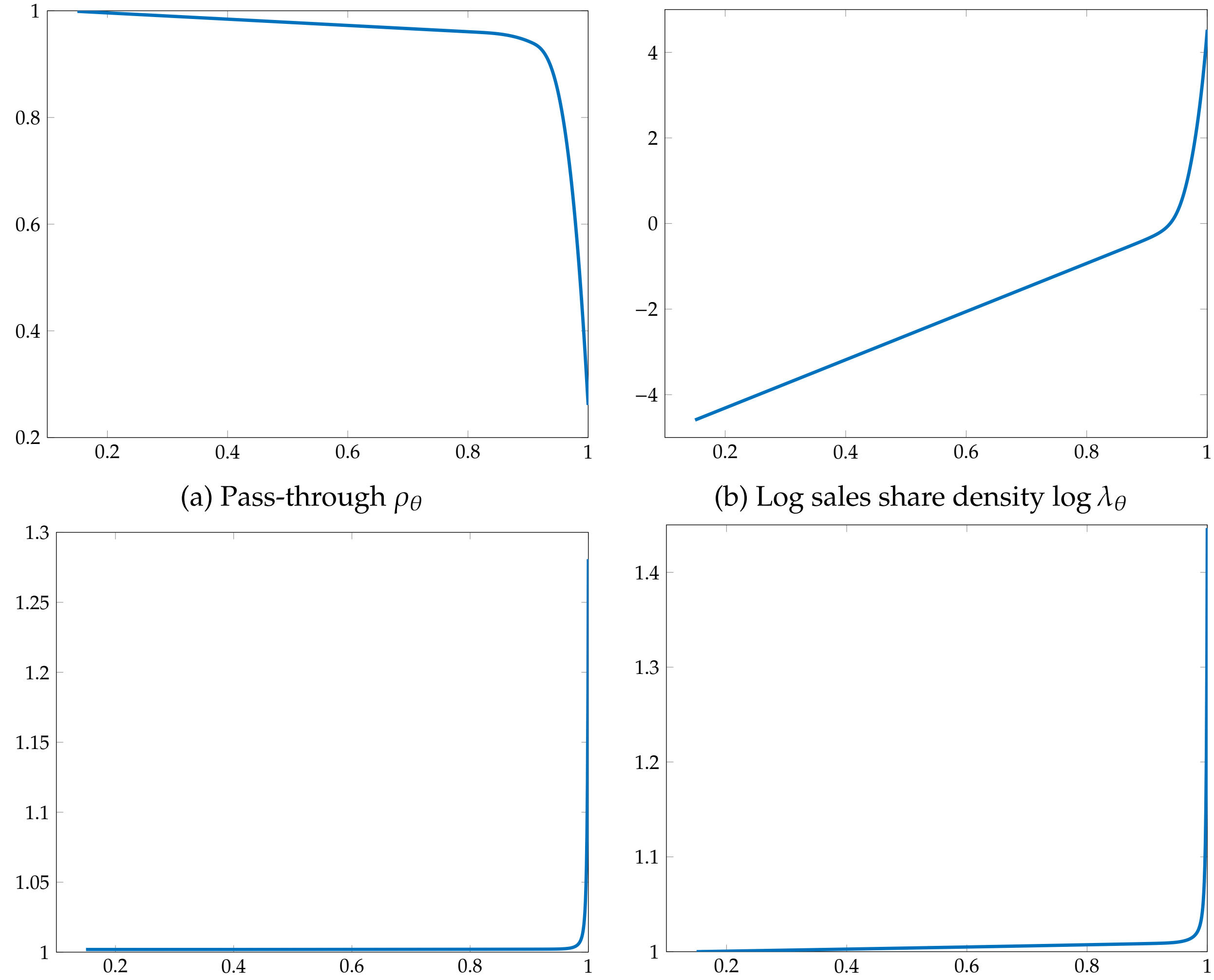
Firms are ranked by increasing size so that the type θ of a given firm is such that $\theta/(1-\theta^*)$ is the number of firms that are smaller. To reflect this parametrization, we use the uniform type density with $g(\theta) = 1$ and $G(\theta) = 1 - \theta$.

Our results require taking a stand on the average markup $\bar{\mu} = 1/[\mathbb{E}_\lambda[1/\mu_\theta]]$ and on the average infra-marginal surplus ratio $\bar{\delta} = \mathbb{E}_\lambda[\delta_\theta]$. To set $\bar{\delta}$, we consider two benchmark calibrations: (1) efficienty entry $\bar{\delta} = \bar{\mu}$, and (2) efficient selection $\bar{\delta} = \delta_{\theta^*}$. We set $\bar{\mu} = 1.045$, which is chosen so that $d \log Y / d \log L = 0.13$ under the assumption that selection is efficient. The number 0.13 is broadly in line with the literature's view about the welfare effects of population changes (e.g. Bartelme et al., 2019). In the Dixit-Stiglitz model, this would correspond to setting an elasticity of substitution around 8.

Estimation results. Figures 3a and 3b display pass-throughs ρ_θ and log sales $\log \lambda_\theta$ as a function of firm type θ. Sales are initially increasing exponentially, but become super-exponential towards the end reflecting a high degree of concentration. Pass-throughs decrease from 1 for the smallest firms to about 0.3 for the largest firms.

Figure 3c shows that markups μ_θ are increasing and convex in θ. The markup ranges from close to zero for the smallest firms to around 30% for the very largest firms. The heterogeneity in markups is a consequence of the vast dispersion in the firm size distribution and estimated pass-throughs. A similar pattern is shown for the productivity/quality shifters of firms in Figure 3d. Figure 4 plots the inverse residual demand curve in linear and log-log terms. Figure 4a shows that our estimate has a distinctly non-isoelastic shape, indicating substantial departures from CES. The Lerner formula ties the markup to the price elasticity of demand σ_θ, which means that σ_θ is around 500 for the very smallest firms, and around 4 for the very largest firms. On the other hand, the log-log plot shows that the residual demand curve is log-concave, which confirms that Marshall's weak second of law of demand holds, and markups are increasing in size.

Figures 5b and 5a show the infra-marginal surplus ratio δ for the efficient-selection case ($\delta_{\theta^*} = \bar{\delta}$) and the efficient-entry case ($\bar{\mu} = \bar{\delta}$). In both cases, δ_θ is U-shaped, although in one case it starts at a much higher level than the other. The fact that δ_θ is non-monotonic means that, contra-Dhingra and Morrow (2019), private and social



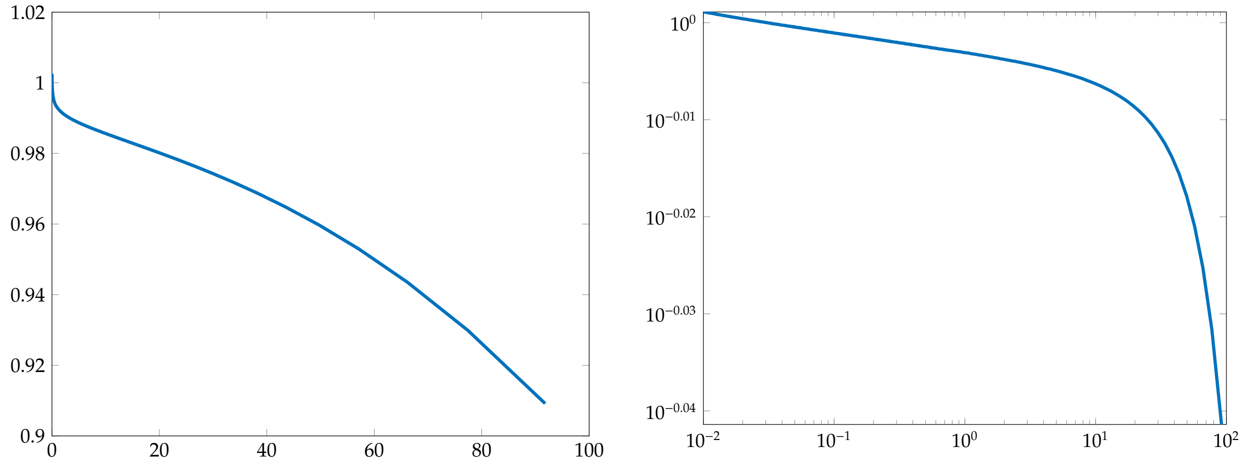
(a) Pass-through ρ_θ

(b) Log sales share density $\log \lambda_\theta$

(c) Markup μ_θ

(d) Productivity A_θ

Figure 3



(a) Inverse residual demand curve

(b) Log-log inverse residual demand curve

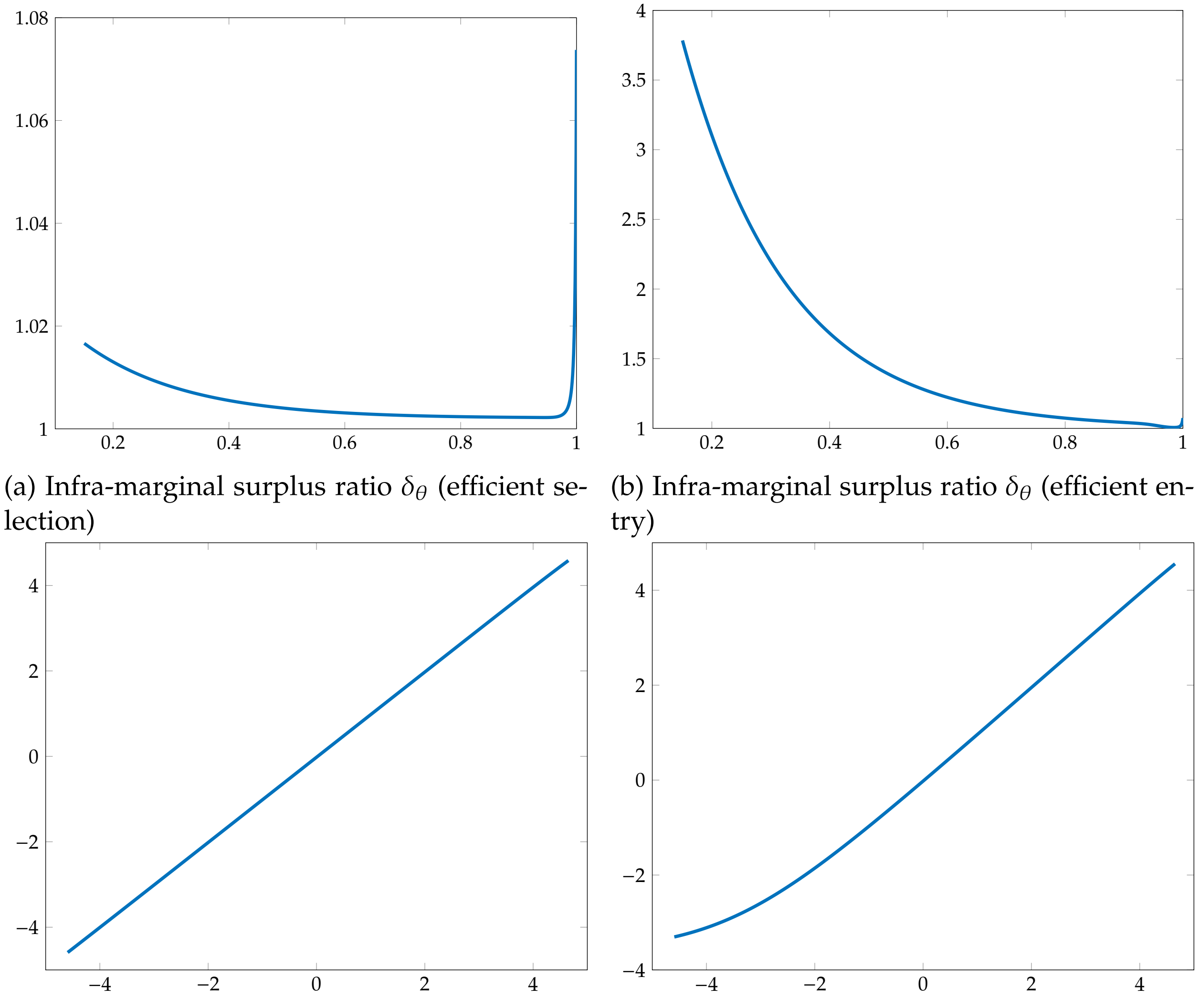
Figure 4: This plots price against quantity for the efficient-entry case. The results for the efficient-selection are similar.

preferences are not globally aligned. Finally, Figures 5c and 5d display the Kimball aggregator for the efficient-selection and efficient-entry case. The fact that the latter is less log-linear at the beginning shows that small firms have higher infra-marginal surplus.

Implications. Since pass-throughs are strictly decreasing, it comes as no surprise that markups are increasing (since Marshall's strong second law holds, the weak second law holds a fortiori). However, since the infra-marginal surplus ratios are U-shaped means that preferences are not globally aligned.

In the efficient entry case, $\delta_{\theta^*} > \bar{\delta}$, meaning that selection is inefficiently too tough. Welfare could be improved by allowing more small firms to operate, and hence, a toughening of the selection cutoff will worsen welfare. Nevertheless, this does not imply that we would want small firms to become larger — since small firms have lower markups, efficiency would be improved by having more small firms, but having them produce less than they already do (since they have relatively low markups). The fact that an increase in the cutoff θ^* reduces welfare may be counterintuitive, since sometimes people argue that an increase in the selection cutoff ipso facto increases efficiency, but this argument is flawed.

On the other hand, in the efficient selection case, $\bar{\delta} < \bar{\mu}$, meaning that there is excessive entry. In both cases, markups μ_θ are increasing in θ indicating that small firms are too large and large firms are too small along the intensive margin.



(a) Infra-marginal surplus ratio δ_θ (efficient selection)

(b) Infra-marginal surplus ratio δ_θ (efficient entry)

(c) Kimball aggregator $\log \Upsilon(\log y/Y)$ (efficient selection)

(d) Residual Demand Curve $\log \Upsilon(\log y/Y)$ (efficient entry)

Figure 5

7.3 Shocks to Population

In this section, we report the elasticities of consumer welfare and real output per capita to changes in population. As stressed before, increases in population can also be interpreted as changes in trade integration.

Baseline. Table 1 implements Proposition 3, reporting the elasticity of consumer welfare to population, its decomposition into pure changes in technology and changes in allocative efficiency

$$\Delta \log Y = \Delta \log Y^{tech} + \Delta \log Y^{alloc}.$$

The table also breaks down the allocative efficiency effect by consider the different margins of adjustment. Welfare under the entry-only allocation Y^{ϵ} (holding fixed θ^* and markups); welfare allowing entry and selection to adjust Y^{ε,θ^*} but holding fixed markups; and welfare when all three margins can adjust Y. Table 1 presents the contributions of endogenous entry, exit, and markups to changes in allocative efficiency. In other words, the sum of the three adjustment rows gives the overall change in allocative efficiency in equilibrium.

Heterogeneous Firms	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$
Welfare: $\Delta \log Y$	0.130	0.145
Technical efficiency: $\Delta \log Y^{tech}$	0.017	0.045
Allocative efficiency: $\Delta \log Y^{alloc}$	0.114	0.100
Adjustment of Entry: $\Delta \log Y^{\epsilon} - \Delta \log Y^{tech}$	0.117	0.408
Adjustment of Exit: $\Delta \log Y^{\epsilon,\theta^*} - \Delta \log Y^{\epsilon}$	0.000	-0.251
Adjusutment of Markups: $\Delta \log Y - \Delta \log Y^{\epsilon,\theta^*}$	-0.004	-0.057
Real GDP per capita	0.024	0.024
Average markup	1.045	1.045

Table 1: Decomposition of welfare effects of a population shock into technical and allocative efficiency following Proposition 3. The decomposition of allocative efficiency into to entry, exit, and markups also follows Proposition 3. The real GDP response follows Proposition 4. Average markup is the harmonic average, set so that the welfare response to population shocks is 0.13.

Symmetric Firms	$\bar{\delta} = \delta_{\theta^*}$	$\bar{\delta} = \bar{\mu}$
Welfare: $\Delta \log Y$	0.019	0.045
Technical efficiency: $\Delta \log Y^{tech}$	0.017	0.045
Allocative efficiency: $\Delta \log Y^{alloc}$	0.002	0.000
Real GDP per capita	0.003	0.003
Average markup	1.045	1.045

Table 2: The elasticity of welfare and real GDP to a population shock following Propositions 1 and 7. The average markup and infra-marginal consumer surplus ratio $\bar{\delta}$ are kept the same as in the corresponding column in Table 1.

We start by discussing the case with entry-efficiency first. By construction, the elasticity of consumer welfare to population is 0.13. Only around a tenth of the overall effect is due to the pure technology effect $\bar{\delta} - 1 = 0.017$. Changes in allocative efficiency 0.114 account for around nine tenths of the overall effect. An Increase in population therefore brings about considerable improvements in allocative efficiency, and these improvements are larger than the pure technology effects arising directly from technological increasing returns.

The changes in allocative efficiency from endogenous entry are large and positive at 0.117. Increases in population lead to a reduction in the aggregate price index for substitution $\bar{\delta}/Y$. This causes a larger reduction in quantity per capita for small firms with high elasticities than for large firms with low elasticities. This reallocation towards large firms, which were too small to begin with, and away from small firms, which were too small to begin with, improves allocative efficiency. The changes in allocative efficiency from endogenous exit is zero by construction since $\bar{\delta} = \delta_{\theta^*}$. Finally, the endogenous changes in markups has a slightly negative effect of -0.004. The reason is subtle, and there are several effects to consider. First, increases in in population lead to a reduction in the price index $\bar{\delta}/Y$. This triggers a pro-competitive effect by causing an overall reduction in markups, and a reduction in entry, which is beneficial since there was too much entry to begin with (the average markup is higher than the average infra-marginal surplus ratio). However, the changes in markups are not uniform, and larger firms cut their markups by more than smaller firms since they have lower pass-through. As before, large firms also face less elastic demand curves than small firms, so that the overall reallocation across large and small firms is, in principle, ambiguous and depends on whether the pass-through effect dominates the elasticity effect. Since the overall effect of the change in markups is negative, we know that this detrimental reallocation effect dominates.

The elasticity of real GDP per capita is small at 0.024. This is much smaller than the elasticity of consumer welfare. This is a consequence of the well-known result that the welfare benefits of new goods are not reflected in changes in real output. Indeed, the positive changes in real output can be entirely attributed to the reduction in markups of existing firms. In particular, it also worth noting as before, that the movement in the productivity cutoff θ^* on its own, holding fixed markups, plays no particular role in determining real GDP or aggregate TFP, even though the model is not efficient. Therefore, an increase in the cutoff does not translate into an improvement in aggregate productivity (as it is measured).

Next, consider the efficient-entry case. The elasticity of welfare with respect to population shocks is now slightly higher at 0.145. The pure technology effect is now 0.045, reflecting the fact that $\bar{\delta}$ is calibrated to equal $\bar{\mu} = 1.045$. The allocative efficiency effect is still much more important than the pure technology effect at 0.100.

The changes in allocative efficiency from endogenous entry are now much larger at 0.408. The intuition is the same as before: increases in population lead to a reduction in the aggregate price index, and this shifts resources away from small firms facing elastic firms towards larger firms facing relatively more inelastic demand. The reason the effects are so much larger than they were in the efficient-selection case is because $E_\lambda(\delta_\theta) - 1$ is now 0.045 instead of 0.017, meaning that entry is more valuable than it was before. Furthermore, since the new entry moves the price index, and the changes in the price index cause large firms to expand relative to small firms, there is a feedback loop from new entry, to changes in the price index, to changes in aggregate profits, back to entry, amplifying the effect.

The exit margin is now non-zero and negative at -0.27. The reason for this can be seen from inspecting Figure 5b, which shows that the infra-marginal surplus at the cutoff is much higher than average, hence, as the cutoff increases in response to toughening competition, very socially valuable small firms are forced to exit.

Finally, the markup effect is still negative and larger in magnitude at -0.057. The reason the markup effect is now more negative than it was before is because in the efficient-selection case there was too much entry, so the overall reductions in markups had a positive effect (over and above the detrimental reallocation across existing firms). Since we are now imposing entry-efficiency, this latter effect no longer operates, and the overall contribution of changing markups is more negative than before.

The response of real GDP per capita is basically unchanged at 0.024, since in both specifications, the average reduction in markups for existing firms is roughly the same.

Homogeneous firms. To emphasize the interaction of heterogeneity and inefficiency, we end this section by comparing our model to a model with homogeneous firms, calibrated to have a pass-through equal to the average (sales-weighted) pass-through and a markup equal to the harmonic average. The results can be found in Table 2.

The most striking difference is that the elasticity of consumer welfare to population is much smaller, because changes in allocative efficiency become negligible. For the efficient-entry specification $\bar{\delta} = \bar{\mu} = 1.045$, the allocative efficiency effects are exactly zero since the model is efficient, and for efficient-selection specification $\bar{\delta} = 1.017$, they are approximately zero. This is because there are no longer changes in allocative efficiency from the entry margin (holding fixed selection and markups) or from the selection margin (holding fixed markups) because there is no heterogeneity. Instead, all the changes in allocative efficiency come from the pro-competitive (markup-reducing) effects of population increases. These much discussed pro-competitive effects are very small in comparison the changes in allocative efficiency arising from reallocations across heterogenous firms that we discussed above. In fact, compared to the benchmark model with heterogeneity, they have the opposite sign. Similar observations apply to response of real GDP per capita.

8 Conclusion

This paper analyzes how changes in market size affect welfare and real GDP in a model with monopolistic competition. We decompose the overall change into changes in technical and allocative efficiency. We use firm-level information to non-parametrically recover preferences and quantify our decomposition. We find that changes in allocative efficiency, due to reallocations of resources, are overwhelmingly the most important source for welfare gains.

References

Amiti, Mary, Oleg Itskhoki, and Jozef Konings, "International Shocks, Variable Markups, and Domestic Prices," *The Review of Economic Studies*, 02 2019.

Arkolakis, Costas, Arnaud Costinot, Dave Donaldson, and Andrés Rodríguez-Clare, "The elusive pro-competitive effects of trade," *The Review of Economic Studies*, 2018, *86* (1), 46–80.

Baqaee, David Rezza and Emmanuel Farhi, "Productivity and Misallocation in General Equilibrium.," Technical Report, National Bureau of Economic Research 2017.

_ **and** _ , "Productivity and Misallocation in General Equilibrium.," Technical Report, National Bureau of Economic Research 2017.

Bartelme, Dominick G., Arnaud Costinot, Dave Donaldson, and Andres Rodriguez-Clare, "The Textbook Case for Industrial Policy: Theory Meets Data," NBER Working Papers 26193, National Bureau of Economic Research, Inc August 2019.

Bilbiie, Florin O, Fabio Ghironi, and Marc J Melitz, "Monopoly power and endogenous product variety: Distortions and remedies," *American Economic Journal: Macroeconomics*, 2019, *11* (4), 140–74.

Chamberlin, Edward Hastings, *Theory of monopolistic competition: A re-orientation of the theory of value*, Oxford University Press, London, 1933.

Dhingra, Swati and John Morrow, "Monopolistic competition and optimum product diversity under firm heterogeneity," *Journal of Political Economy*, 2019, *127* (1), 196–232.

Dixit, Avinash K and Joseph E Stiglitz, "Monopolistic competition and optimum product diversity," *The American economic review*, 1977, *67* (3), 297–308.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu, "How costly are markups?," Technical Report, National Bureau of Economic Research 2018.

Epifani, Paolo and Gino Gancia, "Trade, markup heterogeneity and misallocations," *Journal of International Economics*, 2011, *83* (1), 1–13.

Hopenhayn, Hugo A, "Entry, exit, and firm dynamics in long run equilibrium," *Econometrica: Journal of the Econometric Society*, 1992, pp. 1127–1150.

Hsieh, Chang-Tai and Peter J Klenow, "Misallocation and manufacturing TFP in China and India," *The Quarterly journal of economics*, 2009, *124* (4), 1403–1448.

Hulten, Charles R, "Growth Accounting with Intermediate Inputs," *The Review of Economic Studies*, 1978, pp. 511–518.

Kimball, Miles, "The Quantitative Analytics of the Basic Neomonetarist Model," *Journal of Money, Credit and Banking*, 1995, *27* (4), 1241–77.

Klenow, Peter J and Jonathan L Willis, "Real rigidities and nominal price changes," *Economica*, 2016, *83* (331), 443–472.

Krugman, Paul R, "Increasing returns, monopolistic competition, and international trade," *Journal of international Economics*, 1979, *9* (4), 469–479.

Mankiw, N. Gregory and Michael D. Whinston, "Free Entry and Social Inefficiency," *RAND Journal of Economics*, Spring 1986, *17* (1), 48–58.

Melitz, Marc J., "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica*, November 2003, *71* (6), 1695–1725.

Melitz, Marc J, "Competitive effects of trade: theory and measurement," *Review of World Economics*, 2018, *154* (1), 1–13.

Pugsley, Benjamin W, Petr Sedlacek, and Vincent Sterk, "The nature of firm growth," 2018.

Restuccia, Diego and Richard Rogerson, "Policy distortions and aggregate productivity with heterogeneous establishments," *Review of Economic dynamics*, 2008, *11* (4), 707–720.

Robinson, Joan, *The economics of imperfect competition*, Springer, 1933.

Spence, Michael, "Product selection, fixed costs, and monopolistic competition," *The Review of economic studies*, 1976, *43* (2), 217–235.

Venables, Anthony J, "Trade and trade policy with imperfect competition: The case of identical products and free entry," *Journal of International Economics*, 1985, *19* (1-2), 1–19.

Vives, Xavier, *Oligopoly pricing: old ideas and new tools*, MIT press, 1999.

Zhelobodko, Evgeny, Sergey Kokovin, Mathieu Parenti, and Jacques-François Thisse, "Monopolistic competition: Beyond the constant elasticity of substitution," *Econometrica*, 2012, *80* (6), 2765–2784.

Appendix A Details of Empirical Implementation

Amiti et al. (2019) provide estimates of the average sales-weighted pass-through (denoted by α) for Belgian manufacturing firms conditional on the firms being smaller than a certain size as measured by their numbers of employees. These estimates are based on information from Prodcom, which is a subsample of Belgian manufacturing firms. Inclusion in Prodcom requires that firms have turn-overs above 1 million euros, which means that the sample is not representative of all manufacturers. The estimates are in Table **??**.

No of employees	Share of observations	Share of employment	Share of sales	α
100	0.76313963	0.14761668	0.23096292	0.9719
200	0.85435725	0.22086396	0.3389753	0.8689
300	0.88848094	0.28832632	0.4083223	0.9295
400	0.92032149	0.33549505	0.48074553	0.8303
500	0.93746047	0.38345889	0.54008827	0.6091
600	0.94523549	0.41987701	0.58209142	0.6612
1000	0.96365488	0.52280162	0.66820585	0.6229
8000	0.99996915	0.99999999	0.99999174	0.6497

Table 3: Estimates from Amiti et al. (2019).

Our objective is to infer the pass-through ρ as a function of firm size. With some abuse of notation, let $\theta \in [0, 1]$ be the fraction of observations in Prodcom up to some sales value. Let $\lambda(\theta)$ be the sales share density of Prodcom firms of type θ. Then the variable "Share of sales" is defined as

$$\Lambda(\theta) = \int_0^\theta \lambda(x)dx.$$

We fit a smooth curve to $\Lambda(\theta)$, then the pdf of sales shares $\lambda(\theta)$ is given by

$$\lambda(\theta) = \frac{d\Lambda}{d\theta}.$$

The curve we fit has the form $\exp(c_0 + c_1\theta + c_2\theta^{c_3})$, where c_0, c_1, c_2, c_3 are chosen to minimize the mean squared error.

Next, the variable $\alpha(\theta)$ satisfies

$$\alpha(\theta) = \frac{\int_0^\theta \lambda(x)\rho(x)dx}{\int_0^\theta \lambda(x)dx},$$

$$= \frac{\int_0^\theta \lambda(x)\rho(x)dx}{\Lambda(\theta)},$$

where $\lambda(x)$ is the sales-share of firms of type x. Next we fit a flexible spline function to $\alpha(\theta)$. The fitted curve is shown in Figure 6.

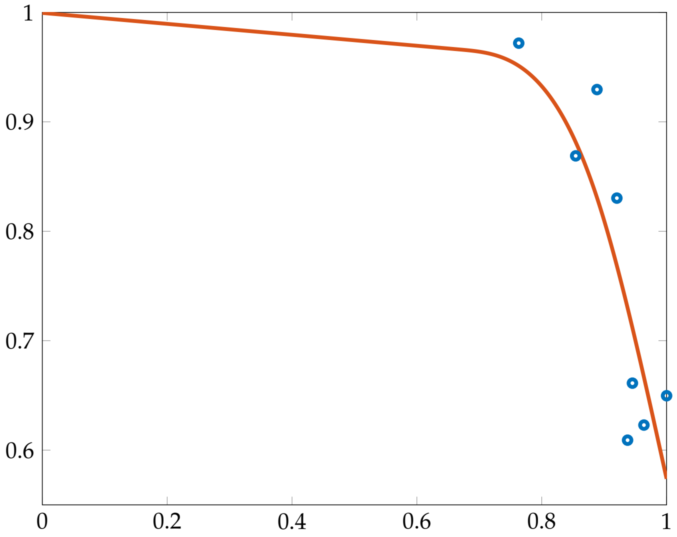


Figure 6: Average pass-through for firms up to a certain size α from Prodcom.

To recover the pass-throughs $\rho(\theta)$, we write

$$\frac{d\alpha}{d\theta} = \frac{\lambda(\theta)\rho(\theta)}{\int_0^\theta \lambda(x)dx} - \frac{\lambda(\theta)}{\int_0^\theta \lambda(x)dx}\alpha(\theta).$$

In other words, we can recover the pass-through function via

$$\begin{aligned}\rho(\theta) &= \frac{\left(\int_0^\theta \lambda(x)dx\right)}{\lambda(\theta)}\frac{d\alpha}{d\theta} + \alpha(\theta), \\ &= \frac{\Lambda(\theta)}{\lambda(\theta)}\frac{d\alpha}{d\theta} + \alpha(\theta).\end{aligned}$$

This gives us pass-throughs as a function of the number of employees.

Next, we use information from VAT declaration in Belgium for the year 2014 to recover the sales distribution of Belgian manufacturers (overcoming the sample selection issues in Prodcom). Table 4 displays the underlying data.

Number of employees	Share of sales	Share of Observations
1	0.004559	0.16668
2	0.00826	0.284539
3	0.014786	0.375336
5	0.022269	0.489659
10	0.043011	0.652879
20	0.076444	0.779734
30	0.111713	0.843161
50	0.163492	0.906204
75	0.198242	0.932729
100	0.231815	0.947413
200	0.325376	0.974629
300	0.386449	0.983547
400	0.449491	0.989237
500	0.486108	0.991927
600	0.655522	0.994311
1000	0.740656	0.997386
8000	0.970654	0.999923

Table 4: Firm size distribution for manufacturing firms from VAT declarations in Belgium for 2014.

As before, we let $\theta \in [0,1]$ index the fraction of observations up to some size. Then the variable "Share of sales" is defined as

$$\Lambda(\theta) = \int_0^\theta \lambda(x) dx,$$

where (abusing notation) λ is the sales share density of all manufacturing firms (rather than just the ones in Prodcom). We fit a smooth curve to $\Lambda(\theta)$, then the pdf of sales shares $\lambda(\theta)$ is given by

$$\lambda(\theta) = \frac{d\Lambda}{d\theta}.$$

The curve we fit has the form $\exp(c_0 + c_1\theta + c_2\theta^{c_3})$, displayed in Figure 7. Finally, we merge our pass-through information from Prodcom with the sales density from VAT declarations by assuming that the pass-through ρ of a firm with a given number of employees in Prodcom is the same as it is in the bigger dataset. We then fit a smooth spline to this pass-through data from $[0,1]$ assuming that the pass-through for the smallest firm is 1 and declines monotonically from the smallest firm to the first observation (which is a pass-through of 0.97 for firms with 100 employees). Given

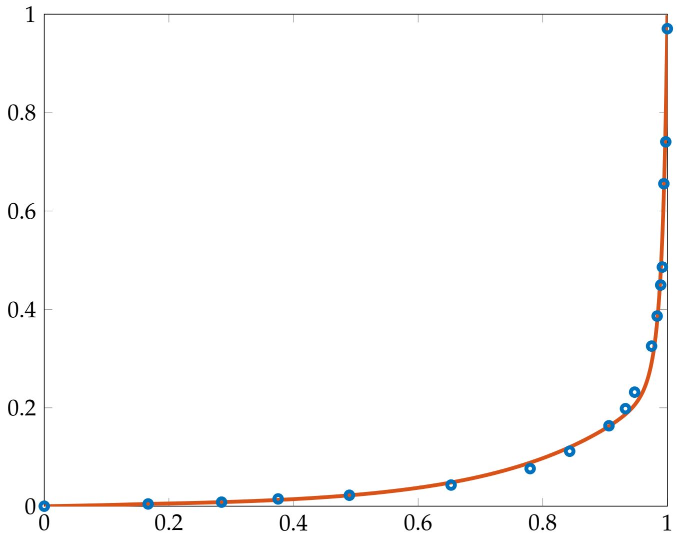


Figure 7: Cumulative share distribution Λ_θ from VAT declarations.

a smooth curve for both λ_θ and ρ_θ we follow the procedure outlined in Section 7.1, solving the differential equations numerically using the Runge-Kutta algorithm on a large grid.

Appendix B Propagation and Aggregation Equations

In this section, we summarize the propagation and aggregation equations for the model with heterogenous firms. We expand the equilibrium equations presented in Section 2.2 to the first order in the shocks. Changes in all the equilibrium variables are expressed via propagation equations as functions of changes in consumer welfare. Changes in consumer welfare are then expressed as as functions of the changes in the equilibrium variable via an aggregation equation. Putting propagation and aggregation together yields a fixed point in changes in consumer welfare.

Aggregate price index. Differentiating the definition of the demand index, we find

$$-d\log(\frac{\bar{\delta}}{Y}) = -(\lambda_{\theta^*}-1)\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + d\log M + d\log Y + \mathbb{E}_\lambda\left[\left(1-\frac{1}{\sigma_\theta}\right)d\log(\frac{y_\theta}{Y})\right].$$

Combining this equation with the equation for quantities and markups, we get

$$-d\log(\frac{\bar{\delta}}{Y}) = -(\lambda_{\theta^*}-1)\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + d\log M + d\log Y + \mathbb{E}_\lambda\left[\rho_\theta(\sigma_\theta-1)\left(d\log A_\theta + d\log(\frac{\bar{\delta}}{Y})\right)\right].$$

Finally, combining with the first equation for entry derived below, we find

$$d\log(\frac{\bar{\delta}}{Y}) \quad = \quad \frac{-d\log Y - \mathbb{E}_{\lambda(1-1/\mu)}\left[(\sigma_\theta - 1)d\log A_\theta\right] + \frac{\frac{f_e\Delta}{L}d\log(\frac{f_e\Delta}{L})+[1-G(\theta^*)]\frac{f}{L}d\log(\frac{f}{L})}{\frac{f_e\Delta}{L}+[1-G(\theta^*)]\frac{f}{L}}}{\mathbb{E}_{\lambda(1-1/\mu)}\left[\sigma_\theta\right]}.$$

This equation for the aggregate price index can be replaced in all the equations below.

Entry. We derive two equations for free entry. The first equation is obtained as follows. Differentiating the free-entry condition, we find

$$\frac{\frac{f_e\Delta}{L}d\log(\frac{f_e\Delta}{L}) + [1-G(\theta^*)]\frac{f}{L}d\log(\frac{f}{L})}{\frac{f_e\Delta}{L} + [1-G(\theta^*)]\frac{f}{L}} + d\log M = \mathbb{E}_{\lambda(1-1/\mu)}\left[d\log\left(\lambda_\theta\left(1-\frac{1}{\mu_\theta}\right)\right)\right].$$

Combining with the equation for variable profit shares, we get

$$\begin{aligned} d\log M = &-\frac{\frac{f_e\Delta}{L}d\log(\frac{f_e\Delta}{L}) + [1-G(\theta^*)]\frac{f}{L}d\log(\frac{f}{L})}{\frac{f_e\Delta}{L} + [1-G(\theta^*)]\frac{f}{L}} + (\lambda_{\theta^*}-1)\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* \\ &+ \mathbb{E}_{\lambda(1-1/\mu)}\left[(\sigma_\theta-1)\left(d\log A_\theta + d\log(\frac{\bar{\delta}}{Y})\right)\right] - \mathbb{E}_\lambda\left[\rho_\theta(\sigma_\theta-1)\left(d\log A_\theta + d\log(\frac{\bar{\delta}}{Y})\right)\right]. \end{aligned}$$

The second equation is obtained by differentiating the demand index

$$d\log M = (\lambda_{\theta^*}-1)\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* - d\log\bar{\delta} - \mathbb{E}_\lambda\left[\rho_\theta(\sigma_\theta-1)\left(d\log A_\theta + d\log(\frac{\bar{\delta}}{Y})\right)\right].$$

Sales shares. Differentiating the sales shares equation, we find

$$d\log\lambda_\theta = d\log M + d\log Y + (\sigma_\theta-1)d\log(\frac{A_\theta}{\mu_\theta}) + \sigma_\theta d\log(\frac{\bar{\delta}}{Y}).$$

Combining with the second equation for entry, we get

$$\begin{aligned} d\log\lambda_\theta = (\lambda_{\theta^*}-1)\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + \rho_\theta(\sigma_\theta-1)\left(d\log A_\theta + d\log(\frac{\bar{\delta}}{Y})\right) \\ - \mathbb{E}_\lambda\left[\rho_\theta(\sigma_\theta-1)\left(d\log A_\theta + d\log(\frac{\bar{\delta}}{Y})\right)\right]. \end{aligned}$$

Markups. Differentiating the markup equation, we get

$$d\log\mu_\theta = (1-\rho_\theta)\left(d\log A_\theta + d\log(\frac{\bar{\delta}}{Y})\right).$$

Variable profit shares. Combining the equations for sales shares and for markups, we get

$$d\log\left(\lambda_\theta\left(1-\frac{1}{\mu_\theta}\right)\right) = (\lambda_{\theta^*}-1)\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + (\sigma_\theta-1)\left(d\log A_\theta + d\log(\frac{\bar{\delta}}{Y})\right)$$
$$-\mathbb{E}_\lambda\left[\rho_\theta(\sigma_\theta-1)\left(d\log A_\theta + d\log(\frac{\bar{\delta}}{Y})\right)\right].$$

Quantities. Differentiating the individual demand function, we find

$$d\log(\frac{y_\theta}{Y}) = \sigma_\theta\left(d\log(\frac{A_\theta}{\mu_\theta}) + d\log(\frac{\bar{\delta}}{Y})\right).$$

Combining with the equation for markups, we get

$$d\log(\frac{y_\theta}{Y}) = \rho_\theta\sigma_\theta\left(d\log A_\theta + d\log(\frac{\bar{\delta}}{Y})\right).$$

Selection. Differentiating the selection condition, we get

$$(\sigma_{\theta^*}-1)\left(\frac{\partial\log A_\theta}{\partial\theta}|_{\theta=\theta^*}\right)d\theta^* = -d\log\left(\lambda_{\theta^*}\left(1-\frac{1}{\mu_\theta}\right)\right) + d\log M + d\log(\frac{f_o}{L}).$$

Combining with the equations for variable profits shares and entry, we get

$$(\sigma_{\theta^*}-1)\left(\frac{\partial\log A_\theta}{\partial\theta}|_{\theta=\theta^*}\right)d\theta^* = -(\sigma_{\theta^*}-1)\left(d\log A_\theta + d\log(\frac{\bar{\delta}}{Y})\right) - d\log\bar{\delta} + d\log(\frac{f_o}{L}),$$

where we note that

$$\frac{\partial\log A_\theta}{\partial\theta}|_{\theta=\theta^*} = \frac{g(\theta^*)}{g_a(\log A_{\theta^*})}.$$

Welfare. Differentiating the consumer welfare equation, we get

$$d\log Y = (\bar{\delta}-\delta_{\theta^*})\lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + (\bar{\delta}-1)\,d\log M + \mathbb{E}_\lambda\left[d\log(\frac{A_\theta}{\mu_\theta})\right].$$

Combining with the equation for markups, we get

$$d\log Y = (\bar{\delta}-\delta_{\theta^*})\lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}d\theta^* + (\bar{\delta}-1)d\log M + \mathbb{E}_\lambda\left[\rho_\theta d\log A_\theta - (1-\rho_\theta)d\log(\frac{\bar{\delta}}{Y})\right].$$

Combining with the equations for the aggregate price index and entry leads to a fixed point in $d\log Y$.

Appendix C Additional Comparative Statics

In this section, we characterize comparative statics with respect to shocks to the fixed costs and shocks to the productivity distribution. We start with fixed cost shocks, and then examine productivity shocks.

C.1 Shocks to Fixed Costs

As with population shocks, we begin by focusing on the homogeneous firms case.

C.1.1 Homogeneous Firms

Proposition 6. *Suppose that firms have the same productivity $A_\theta = A$. In response to changes in population $d\log L$, changes in consumer welfare are given by*

$$d\log Y = \underbrace{(\delta-1)d\log L}_{\textit{pure technology}} + \underbrace{\delta\frac{\xi}{1-\xi}d\log L}_{\textit{allocative efficiency}},$$

where

$$\xi = \left(1-\rho\right)\left(1-\frac{\delta-1}{\mu-1}\right)\frac{1}{\sigma} = \left(1-\rho\right)\left(1-\frac{\delta}{\mu}\right).$$

Changes in entry costs $d\log(f_e\Delta)$ and in overhead costs $d\log f_0$ respectively have the same effects on consumer welfare as change in population shocks $d\log L = -[f_e\Delta/(f_e\Delta+f_0)]d\log(f_e\Delta)$ and $d\log L = -[f_o/(f_e\Delta+f_0)]d\log(f_o)$.

Proposition 7. *Suppose that firms have the same productivity $A_\theta = A$. In response to changes in population $d\log L$, changes in real output per capita are given by*

$$d\log Q^p = \frac{1-\rho}{\sigma}(d\log Y + d\log L),$$

where $d\log Y$ is given by Proposition 1. Changes in entry costs $d\log(f_e\Delta)$ and in overhead costs $d\log f_0$ respectively have the same effects on these variables as change in population shocks $d\log L = -[f_e\Delta/(f_e\Delta + f_o)]d\log(f_e\Delta)$ and $d\log L = -[f_o/(f_e\Delta + f_o)]d\log(f_o)$.

C.1.2 Heterogeneous Firms

Now, we consider shocks to fixed costs when firms are heterogeneous.

Proposition 8. *In response to changes in fixed costs of entry $d\log(f_e\Delta)$ and fixed overhead costs $d\log f_o$, changes in consumer welfare are given by*

$$d\log Y = \underbrace{-\left(\mathbb{E}_\lambda[\delta_\theta] - 1\right)\frac{f_e\Delta d\log(f_e\Delta) + f_o d\log f_o}{f_e\Delta + (1 - G(\theta^*))f_o}}_{\text{pure technology}}$$
$$\underbrace{-\frac{\xi^\epsilon + \xi^\mu + \xi^{\theta^*}}{1 - \xi^\epsilon + \xi^\mu + \xi^{\theta^*}}\left(\mathbb{E}_\lambda[\delta_\theta]\right)\frac{f_e\Delta d\log(f_e\Delta) + (1 - G(\theta^*))f_o d\log f_o}{f_e\Delta + (1 - G(\theta^*))f_o}}_{\text{allocative efficiency}}$$
$$\underbrace{-\frac{\zeta^{\theta^*}}{1 - (\xi^\epsilon + \xi^\mu + \xi^{\theta^*})}\frac{f_e\Delta[d\log(f_e\Delta) - d\log f]}{f_e\Delta + (1 - G(\theta^*))f}}_{\text{allocative efficiency}},$$

where ξ^ϵ, ξ^{θ^*}, and ξ^μ are given in Proposition 3 and

$$\zeta^{\theta^*} = \left(\mathbb{E}_\lambda[\delta_\theta] - \delta_{\theta^*}\right)\left(\lambda_{\theta^*}\gamma_{\theta^*}\frac{1}{\sigma_{\theta^*} - 1}\right).$$

As with population shocks, we can provide sufficient conditions under which changes in allocative efficiency amplify or mitigate the effects of the shocks.

Corollary 3. *Sufficient conditions for positive changes in allocative efficiency in response to decreases in the fixed cost of entry are the same as in Corollary 2. Indeed, (1), (2), and (3) imply $\xi^\epsilon > 0$, $\xi^{\theta^*} > 0$, and $\xi^\mu > 0$. Furthermore, (1) and (3) imply $\zeta^{\theta^*} > 0$. Sufficient conditions for positive changes in allocative efficiency in response to decreases in the fixed overhead cost if selection decreases ($d\theta^* < 0$) are that (1) and (2) hold but that (3) fail (too much selection).*

To understand these results, it is useful to observe that the model is homogeneous of degree zero in fixed costs and population $f_e\Delta$, f, and L. This is because they only matter through fixed costs per capita $(f_e\Delta)/L$ and f/L. This means that joint proportional reductions in fixed costs of entry and fixed overhead costs $d\log(f_e\Delta) = d\log f < 0$ have exactly the same effects on consumer welfare as equivalent increases in population $d\log L = -d\log(f_e\Delta) = -d\log f > 0$.

With homogeneous firms, shocks to fixed costs act like scaled population shocks even in isolation. The equivalent shock to population is inversely related to the shock to the total fixed cost $-[f_e\Delta d\log(f_e\Delta) + (1 - G(\theta^*))f d\log f]/[f_e\Delta + (1 - G(\theta^*))f]$. This is no longer true with heterogenous firms because the two fixed costs impact selection in different ways.

Consider first a reduction in the fixed cost of entry $d\log(f_e\Delta) < 0$. This reduces the total (entry and overhead) fixed cost per entering variety in proportion to the share of the fixed cost of entry in the total fixed cost $[(f_e\Delta)/[f_e\Delta + (1 - G(\theta^*))f]]d\log(f_e\Delta) < 0$. This reduction in fixed cost acts like an equivalent increase in population coupled with an equivalent increase in the fixed overhead cost. The effect of the former was analyzed in Proposition 3 and Corollary 2. The effect of the latter is to further increase the sales shares of exiting varieties by $-[\lambda_{\theta^*}\gamma_{\theta^*}/(\sigma_{\theta^*}-1)][(f_e\Delta)/[f_e\Delta+(1-G(\theta^*))f]]d\log(f_e\Delta) > 0$. This in turn increases consumer welfare by $-[(\mathbb{E}[\delta_\theta]-\delta_{\theta^*})\lambda_{\theta^*}\gamma_{\theta^*}/(\sigma_{\theta^*}-1)][(f_e\Delta)/[f_e\Delta + (1 - G(\theta^*))f]]d\log(f_e\Delta) > 0$ as long as there is too little selection ($\mathbb{E}_\lambda[\delta_\theta] > \delta_{\theta^*}$). The result in the proposition is obtained by solving the fixed point in $d\log Y$.

Consider now a reduction in the fixed overhead cost $d\log f < 0$. The effect on the selection cutoff is reversed compared to the case of a reduction in the fixed cost of entry: compared to an increase in population by $-[(1-G(\theta^*))f/[f_e\Delta+(1-G(\theta^*))f]]d\log(f) > 0$, the increase in the fixed overhead cost reduces the selection cutoff, which typically overcomes the increase in selection associated with the equivalent increase in population. If this is the case, the overall change in consumer welfare from the change in selection is positive if and only if there is too much selection ($\mathbb{E}_\lambda[\delta_\theta] < \delta_{\theta^*}$).

In both cases, and exactly as for population shocks, we can decompose the general equilibrium response by analyzing three successive equilibrium allocations which allow firms to adjust along more and more margins: entry, entry and exit, and entry, exit and markups. All three equilibrium allocations feature the same pure technology effect, but different changes in allocative efficiency, driven by different changes in the allocation of resources. The corresponding changes in consumer welfare are respectively given by Proposition 8, but with $\xi^\mu = \xi^{\theta^*} = 0$ and $\zeta^{\theta^*} = 0$, $\xi^\mu = 0$, and without any modification.

We can also perform the same decomposition for changes in real GDP per capita.

Proposition 9. *In response to changes in fixed costs of entry $d\log(f_e\Delta)$ and fixed overhead* *costs $d\log f$, changes in real GDP per capita are given by*

$$d\log Q = \left(\mathbb{E}_\lambda\left[(1-\rho_\theta)\right]\right)\left(\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\right)\left(d\log Y + \frac{f_e\Delta d\log(f_e\Delta) + (1-G(\theta^*))fd\log f}{f_e\Delta + (1-G(\theta^*))f}\right),$$

where $d\log Y$ is given by Proposition 8.

Proposition 9 can be used to decompose real output per capita along the same lines as the decomposition of welfare in Proposition 8. Setting $\xi^\mu = \xi^{\theta^*} = 0$, $\zeta^{\theta^*} = 0$ and $\rho_\theta = 1$ holds fixed markups and selection but allows entry, setting $\xi^\mu = 0$ and $\rho_\theta = 1$ holds fixed markups but allows entry and selection to adjust, and finally apply Proposition 9 without any modification allows all margins to adjust.

C.2 CES Example

The CES case is once again very simple. We have $\sigma_\theta = \sigma$, $\mu_\theta = \mu = 1/(1-1/\sigma)$, $\rho = 1$, and $\delta = \sigma/(\sigma-1)$. This implies that $\xi^\epsilon = \xi^{\theta^*} = \xi^\mu = 0$. The simplicity of this expression is a consequence of the fact that the equilibrium is efficient.

C.3 Shocks to Productivity

Now, we consider shocks to the distribution of productivity shifters, starting with the homogeneous firm case before moving onto the heterogeneous case.

C.3.1 Homogeneous Firm Case

Whereas the model is not homothetic in population and fixed costs L, f_e, and f, it is homothetic in productivity A.

Proposition 10. *Suppose that firms have the same productivity $A_\theta = A$. In response to changes in productivity $d\log A$, changes in consumer welfare are given by*

$$d\log Y = \underbrace{d\log A}_{\textit{pure technology}} + \underbrace{0}_{\textit{allocative efficiency}}.$$

In response to a positive productivity shock $d\log A > 0$, individual quantities and consumer welfare all increase proportionately with the shock $d\log y = d\log Y = d\log A$. As a result, there is no change in markups $d\log\mu = 0$, and hence individual prices decrease proportionately with the shock $d\log p = -d\log A$. Entry remains unchanged $d\log M = 0$. More generally the allocation of resources actually stays unchanged, that is, the fractions of labor allocated to entry, overhead, and variable production remain unchanged. The absence of reallocations in turn implies that there are no changes in allocative efficiency. There are only pure changes in technology.

Proposition 11. *Suppose that firms have the same productivity $A_\theta = A$. In response to changes in productivity $d\log A$, changes in real output per capita are given by*

$$d\log Q^q = \underbrace{d\log A}_{\text{pure technology}} + \underbrace{0}_{\text{allocative efficiency}},$$

and

$$d\log Q^p = d\log A.$$

Changes in real output per capita measured with prices are given by $d\log Q^p = -d\log p = d\log A - d\log \mu = d\log A$. Basically, the price of each variety is reduced by the amount of the productivity shock, with no change in markups. Changes in real output per capita measured with quantities are given by $d\log Q^q = d\log y = d\log A$. Basically, the per-capita quantity of each variety is increased by the amount of the productivity shock.

C.3.2 Heterogeneous Firm Case

Finally, we consider shocks to productivities when the firm-size distribution is heterogeneous.

Proposition 12. *In response to changes in productivity $d\log A_\theta$, changes in consumer welfare are given by*

$$d\log Y = \underbrace{\mathbb{E}_\lambda\left[d\log A_\theta\right]}_{\text{pure technology}} + \underbrace{\frac{\nu^\epsilon\left[d\log A_\theta\right] + \nu^{\theta^*}\left[d\log A_\theta\right] + \nu^\mu\left[d\log A_\theta\right]}{1-(\xi^\epsilon + \xi^\mu + \xi^{\theta^*})}}_{\text{allocative efficiency}}$$
$$+ \underbrace{\frac{\xi^\epsilon + \xi^\mu + \xi^{\theta^*}}{1-\xi^\epsilon - \xi^\mu - \xi^{\theta^*}}\left(\mathbb{E}_{\lambda(1-1/\mu)}\left[(\sigma_\theta - 1)d\log A_\theta\right] + \mathbb{E}_\lambda\left[d\log A_\theta\right]\right)}_{\text{allocative efficiency}},$$

where ξ^ϵ, ξ^{θ^*}, and ξ^μ are given in Proposition 3 and

$$\nu^\epsilon\left[d\log A_\theta\right] = \left(\mathbb{E}_\lambda\left[\delta_\theta\right] - 1\right)\left(\mathbb{E}_{\lambda(1-1/\mu)}\left[(\sigma_\theta - 1)d\log A_\theta\right] - \mathbb{E}_\lambda\left[(\sigma_\theta - 1)d\log A_\theta\right]\right),$$

$$\nu^{\theta^*}\left[d\log A_\theta\right] = -\left(\mathbb{E}_\lambda\left[\delta_\theta\right] - \delta_{\theta^*}\right)\left(\lambda_{\theta^*}\gamma_{\theta^*}\frac{\sigma_{\theta^*}d\log A_{\theta^*} - \mathbb{E}_{\lambda(1-1/\mu)}\left[\sigma_\theta d\log A_\theta\right]}{\sigma_{\theta^*}-1}\right),$$
$$\nu^{\mu}\left[d\log A_\theta\right] = -\left(\mathbb{E}_\lambda\left[(1-\rho_\theta)\left[1 - \frac{\mathbb{E}_\lambda\left[\delta_\theta\right]-1}{\mu_\theta - 1}\right]d\log A_\theta\right]\right).$$

Exactly as for shocks to population and to fixed costs, we can decompose the general equilibrium response by analyzing three successive equilibrium allocations which allow firms to adjust along more and more margins: entry, entry and exit, and entry, exit and markups. All three equilibrium allocations feature the same pure technology effect given by the sales-weighted changes in productivities, exactly as in Hulten's theorem (Hulten, 1978). These three equilibrium allocations feature different changes in allocative efficiency, driven by different changes in the allocation of resources. The corresponding changes in consumer welfare are respectively given by Proposition 12, but with $\xi^\mu = \xi^{\theta^*} = 0$ and $\nu^\mu[d\log A_\theta] = \nu^{\theta^*}[d\log A_\theta] = 0$, $\xi^\mu = 0$ and $\nu^\mu[d\log A_\theta] = 0$, and without any modification.

Changes in allocative efficiency are given by the sum of two sets of terms. The first set of terms $\nu^\epsilon\left[d\log A_\theta\right]$, $\nu^{\theta^*}\left[d\log A_\theta\right]$, and $\nu^\mu\left[d\log A_\theta\right]$ captures the effects of changes in productivities $d\log A_\theta$ holding the aggregate price index $\bar{\delta}/Y$ constant. The second set of terms capture the effects of changes in the aggregate price index $d\log(\bar{\delta}/Y) = (\mathbb{E}_{\lambda(1-1/\mu)}[(\sigma_\theta - 1)d\log A_\theta] + d\log Y)\mathbb{E}_\lambda[1/\sigma_\theta]$.

We have already discussed the effects of changes in the aggregate price index, for example in Section 5.2. We therefore focus our discussion on the effects of changes in productivities holding the aggregate price index constant. We quickly discuss the intuition for the terms $\nu^\epsilon\left[d\log A_\theta\right]$, $\nu^{\theta^*}\left[d\log A_\theta\right]$, and $\nu^\mu\left[d\log A_\theta\right]$. These terms are then amplified by a multiplier $1/[1-(\xi^\epsilon + \xi^\mu + \xi^{\theta^*})]$ arising from solving the fixed point in $d\log Y$.

The intuition for the term $\nu^\epsilon\left[d\log A_\theta\right]$ is the following. Productivity shocks change prices for given markups, exit behavior, and aggregate price index. The sales shares of varieties with high markups tend to increase if they experience sufficiently higher relative productivity shocks to offset their relatively lower elasticities. If they do, the variable profit share increases, which increases entry by $\mathbb{E}_{\lambda(1-1/\mu)}[(\sigma_\theta - 1)d\log A_\theta] - \mathbb{E}_\lambda[(\sigma_\theta - 1)d\log A_\theta]$ and welfare by $(\mathbb{E}_\lambda[\delta_\theta] - 1)(\mathbb{E}_{\lambda(1-1/\mu)}[(\sigma_\theta - 1)d\log A_\theta] - \mathbb{E}_\lambda[(\sigma_\theta - 1)d\log A_\theta])$.

The intuition for the term $\nu^{\theta^*}\left[d\log A_\theta\right]$ is the following. Productivity shocks change exit behavior for given markups and aggregate price index. The selection cutoff tends to decrease if the productivity increases relatively more and if the elasticity of substitution is relatively higher at the cutoff. If they do does, the sales share of exiting varieties decreases by $(\sigma_{\theta^*} d\log A_{\theta^*} - \mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta d\log A_\theta])/(\sigma_{\theta^*} - 1)$, which changes welfare by $-(\mathbb{E}_\lambda[\delta_\theta] - \delta_{\theta^*})(\sigma_{\theta^*} d\log A_{\theta^*} - \mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta d\log A_\theta])/(\sigma_{\theta^*} - 1)$.

The intuition for the term $\nu^\mu\left[d\log A_\theta\right]$ is the following. Productivity shocks lead to changes in markups for a given aggregate price index. Increases in productivity lead to increases in markups, which increases the variable profit share. This in turn increases entry and changes welfare by $-\mathbb{E}_\lambda[(1-\rho_\theta)[1-[(\mathbb{E}_\lambda[\delta_\theta]-1)/(\mu_\theta-1)]d\log A_\theta]]$.

Signing the overall changes in allocative efficiency is difficult because of offsetting effects. For example if all productivity shocks are identical $d\log A_\theta = d\log A$, then there are no changes in allocative efficiency, since just like in the case with homogeneous firms, the model is homothetic with respect to such shocks. In this special case, the terms capturing the effects of changes in productivities given the aggregate price index exactly offset (term by term) the terms capturing the effects of changes in the aggregate price index given productivities: the terms in $\nu^\epsilon[d\log A_\theta]$ exactly offset the terms in ξ^ϵ, the terms in $\nu^{\theta^*}[d\log A_\theta]$ exactly offset the terms in ξ^{θ^*}, and the terms in $\nu^\mu[d\log A_\theta]$ exactly offset the terms in ξ^μ. This shows that changes in allocative efficiency from productivity shocks depend finely on the distribution of these shocks across types.

It turns out to be easier to determine if changes in consumer welfare are greater than sales- and pass-through-weighted changes in productivity.

Corollary 4. *Sufficient conditions for changes in consumer welfare to be greater than sales- and pass-through-weighted changes in productivity*

$$d\log Y > \mathbb{E}_\lambda\left[\rho_\theta d\log A_\theta\right]$$

in response to positive changes in productivity are the conditions (1), (2), and (3) of Corollary 2, together with two conditions ensuring that productivity shocks are sufficiently skewed towards large firms

$$\mathbb{E}_{\lambda(1-1/\mu)}\left[(\sigma_\theta - 1)d\log A_\theta\right] - \mathbb{E}_\lambda\left[\rho_\theta(\sigma_\theta - 1)d\log A_\theta\right] > 0,$$

and

$$\mathbb{E}_{\lambda(1-1/\mu)}\left[(\sigma_\theta - 1)d\log A_\theta\right] - (\sigma_{\theta^*} - 1)d\log A_{\theta^*} > 0.$$

Finally, we can apply the same decomposition as above into three different equilibrium allocations incorporating more and more margins of adjustment: entry, entry and exit, and entry, exit and markups. The corresponding changes in real output per capita are respectively given by Proposition 13 below, but with $\xi^\mu = \xi^{\theta^*} = 0$ and $\nu^\mu[d\log A_\theta] = \nu^{\theta^*}[d\log A_\theta] = 0$ and $\rho_\theta = 1$, $\xi^\mu = 0$ and $\nu^\mu[d\log A_\theta] = 0$ and $\rho_\theta = 1$, and without any modification.

Proposition 13. *In response to changes in productivities $d\log A_\theta$, changes in real output per capita are given by*

$$
\begin{aligned}
d\log Q^q = {} & \underbrace{\mathbb{E}_\lambda\left[d\log A_\theta\right]}_{\textit{pure technology}} - \underbrace{\mathbb{E}_\lambda\left[(1-\rho_\theta)\sigma_\theta d\log A_\theta\right]}_{\textit{allocative efficiency}} \\
& + \underbrace{\left(\mathbb{E}_\lambda\left[(\sigma_\theta-1)d\log A_\theta\right] - \mathbb{E}_{\lambda(1-1\mu)}\left[(\sigma_\theta-1)d\log A_\theta\right]\right)}_{\textit{allocative efficiency}} \\
& + \underbrace{\left(1-\mathbb{E}_\lambda\left[\rho_\theta\sigma_\theta\right]\right)\left(\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\right)\left(d\log Y + \mathbb{E}_{\lambda(1-1\mu)}\left[(\sigma_\theta-1)d\log A_\theta\right]\right)}_{\textit{allocative efficiency}},
\end{aligned}
$$

and

$$
\begin{aligned}
d\log Q^p = {} & \mathbb{E}_\lambda\left[\rho_\theta d\log A_\theta\right] \\
& + \left(\mathbb{E}_\lambda\left[(1-\rho_\theta)\right]\right)\left(\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\right)\left(d\log Y + \mathbb{E}_{\lambda(1-1\mu)}\left[(\sigma_\theta-1)d\log A_\theta\right]\right),
\end{aligned}
$$

where $d\log Y$ is given by Proposition 12.

Appendix D Differences in Tastes and Overhead Costs

In this section, we extend the model to allow for differences in tastes and overhead costs, by allowing the Kimball aggregator $\Upsilon(\frac{\theta}{Y};\theta)$ and the overhead cost $f_o(\theta)$ to depend on the type θ of the variety. Instead of ranking types by productivity, we rank them in increasing order of variable profits to overhead cost ratio so that $X_\theta = \lambda_\theta(1-1/\mu_\theta)/f_{o,\theta}$ is increasing in θ. The formulas in the paper continue to apply, with one exception: changes in selection are now given by

$$
\left(\frac{\partial \log X_\theta}{\partial\theta}|_{\theta=\theta^*}\right)d\theta^* = -(\sigma_{\theta^*}-1)\left(d\log A_\theta - d\log(\frac{\bar{\delta}}{Y})\right) - d\log\bar{\delta} + d\log(\frac{f_{o,\theta^*}}{L}).
$$

This implies that in all the formulas, we must now use $\gamma^*_\theta = [g_x(\log X_{\theta^*})/[1-G_x(\log X_{\theta^*})]]/(\sigma_{\theta^*}-1)$ where $g_x(\log X_\theta) = g(\theta)/(\partial\log X_\theta/\partial\theta)$.

Empirical implementation requires more data than the strategy described in Section 7.1. This is because the model is richer. To simplify the discussion, assume that overhead costs are homogeneous so that $f_{o,\theta} = f_o$.

The model without taste shocks required data on sales λ_θ and pass throughs ρ_θ as well as taking a stand on the average markup $\bar{\mu} = 1/[\mathbb{E}_\lambda[1/\mu_\theta]]$ and the average infra-marginal surplus ratio $\bar{\delta} = \mathbb{E}_\lambda[\delta_\theta]$. The nonlinear model could then be perfectly identified, allowing us to perform local and global counterfactuals.

Identification of the model with taste shocks requires additional data: we need data on markups μ_θ and we need to take a stand on the whole distribution of infra-marginal consumption surplus ratios δ_θ. Even with this data, we only have a local identification of the model, allowing us only to perform local first-order counterfactuals.

The reason is that in the model without taste shocks, a bigger firm is a smaller firm which received a positive productivity shock. Cross-sectional observations then allow us to trace the whole individual demand curve and hence to back out the Kimball aggregator up to some constants. This simplification disappears in the model with taste shocks.

Appendix E Real GDP

In a neoclassical setting (without non-convexities), real GDP can in principle be measured in two equivalent ways, either using a Divisia quantity index or a Divisia price index. In this model, since new goods enter with finite sales, this breaks the equivalence between the two indices. The price index is the definition we adopt in the body of the paper, however, for completeness, we also discuss the quantity index. The quantity index measures the change in individual quantities at constant prices

$$d \log Q^q = \mathbb{E}_\lambda[d \log y_\theta].$$

This is equal to

$$d \log Q^q = -d \log M + \lambda_{\theta^*} \frac{g(\theta^*)}{1 - G(\theta^*)} d\theta^* + \mathbb{E}_\lambda \left[d \log(\frac{A_\theta}{\mu_\theta}) \right],$$

The two notions of changes in real output per capita differ. For the rest of this section, denote the price-index notion (that we use in the body of the paper) using $\mathrm{d} \log Q^p$: this is the change in real GDP per capita measured at constant quantities (more precisely, the price index is measured at constant quantities, and then changes in real GDP are defined to be changes in nominal GDP deflated by the price index). Changes in real output per capita measured with quantities $d \log Q^p$ depend only on changes in prices $d \log(p_\theta/w) = d \log(\mu_\theta/A_\theta)$. For given prices $p_\theta/w = \mu_\theta/A_\theta$, they do not depend on the allocation of spending between new, existing, and disappearing varieties. By contrast, changes in real output measured with quantities do depend on the allocation spending for given prices. In fact, $d\log Q^q$ penalizes new product creation since the quantity of new products produced is not included in the measure, but the reduction in the quantity of existing products is included. The reduction in the quantity of existing products comes about from the fact that, in order to produce new products, less of the old products must be produced.

Since real GDP measured at constant prices has a physical interpretation, we can write real output per capita measured with quantities $\mathcal{Q}^q(\mathcal{A},\mathcal{X})$. However, no such representation is available for real output measured with prices Q^p. and

$$d\log Q^q = \underbrace{\frac{\partial \log \mathcal{Q}^q}{\partial \log \mathcal{A}} d\log \mathcal{A}}_{\text{pure technology}} + \underbrace{\frac{\partial \log \mathcal{Q}^q}{\partial \mathcal{X}} d\mathcal{X}}_{\text{allocative efficiency}} .$$

Note that changes in allocative efficiency are different for consumer welfare $d\log Y$ and for changes in real output per capita at constant prices $d\log Q^q$. Changes in allocative efficiency are changes in the object of interest originating in reallocation effects. It is therefore natural that they depend on the object of interest.

Homogeneous Firms

Proposition 14. *Suppose that firms have the same productivity $A_\theta = A$. In response to changes in population $d\log L$, changes in real output per capita are given by*

$$d\log Q^q = \underbrace{-d\log L}_{\textit{pure technology}} + \underbrace{(1-\rho)(d\log Y + d\log L)}_{\textit{allocative efficiency}},$$

and

$$d\log Q^p = \frac{1-\rho}{\sigma}(d\log Y + d\log L),$$

where $d\log Y$ is given by Proposition 1. Changes in entry costs $d\log(f_e\Delta)$ and in overhead costs $d\log f_0$ respectively have the same effects on these variables as change in population shocks $d\log L = -[f_e\Delta/(f_e\Delta + f_o)]d\log(f_e\Delta)$ and $d\log L = -[f_o/(f_e\Delta + f_o)]d\log(f_o)$.

Changes in real output per capita measured with quantities are given by $d\log Q^q = d\log y$ so that $d\log Q^q = d\log Y + d\log(y/Y) = d\log Y - \rho(d\log Y + d\log L)$. They can be decomposed into pure changes in technology $-d\log L$ and changes in allocative efficiency $(1-\rho)d\log Y + (1-\rho)d\log L$.

Holding the allocation of resources constant, an increase in population $d\log L > 0$ leads to a proportional reduction $-d\log L < 0$ in the per-capita quantity of each variety because the number of varieties increases by $d\log L > 0$. The new varieties do not contribute at all to changes in real output measured with quantities. This explains, in this case, the negative pure changes in technology $-d\log L < 0$.

Turning to changes in allocative efficiency, the pro-competitive reduction in markups reduces entry and increases the per-capita quantity of each variety. This explains, in this case, the positive changes in allocative efficiency $(1-\rho)(d\log Y + d\log L) > 0$.

CES Example Changes in real output per capita are given by

$$d\log Q^q = \underbrace{-d\log L}_{\text{pure technology}} + \underbrace{0}_{\text{allocative efficiency}}.$$

Even though the CES model is efficient, and there are no changes in allocative efficiency, increases in population reduce real GDP measured using the quantity index. Intuitively, the production of new goods means that fewer units of existing goods are produced per capita. Since the quantity index only measures changes in the quantity of existing goods per capita, it falls in response to the shock.

Heterogeneous Firms

Proposition 15. *In response to changes in population $d\log L$, changes in real output per capita are*

$$d\log Q^q = \underbrace{-d\log L}_{\textit{pure technology}} + \underbrace{\left(1 - \mathbb{E}_\lambda\left[\rho_\theta\sigma_\theta\right]\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\right)\left(d\log Y + d\log L\right)}_{\textit{allocative efficiency}},$$

and

$$d\log Q^p = \left(\mathbb{E}_\lambda\left[(1-\rho_\theta)\right]\right)\left(\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\right)\left(d\log Y + d\log L\right),$$

where $d\log Y$ is given by Proposition 3.

We can apply the same decomposition as above into three different equilibrium allocations incorporating more and more margins of adjustment: entry, entry and exit, and entry, exit and pricing/markups. The corresponding changes in real output per capita are respectively given by Proposition 4, but setting $\xi^\mu = \xi^{\theta^*} = 0$ and $\rho_\theta = 1$ (which holds fixed markups and the cutoffs), $\xi^\mu = 0$ and $\rho_\theta = 1$ (which holds fixed markups but allows the cutoff to adjust), and without any modification (allowing all margins to adjust).

For changes in real output per capita, it is actually even more interesting to study this decomposition in reverse order, because of the more central role played by pricing/markups in the evolution of these variables. This means incorporating more and more margins of adjustment as follows: pricing/markups, pricing/markups and exit, and pricing/markups, entry and exit. The corresponding changes in real output per capita are respectively given by Proposition 4, but with $\xi^{\epsilon} = \xi^{\theta^*} = 0$, $\xi^{\epsilon} = 0$, and without any modification. For example, under assumptions (1), (2), and (3), changes in real output per capita measured with prices increase as more and more margins of adjustment are incorporated.

Appendix F Closed-Form Solution to Differential Equations

This section provides closed-form solutions to the differential equations defined in Section 7.1. Starting with the differential equation for markups, we have

$$\frac{d\log\mu_\theta}{d\theta} = (\mu_\theta - 1)\frac{1-\rho_\theta}{\rho_\theta}\frac{d\log\lambda_\theta}{d\theta}.$$

We can rewrite this as

$$\frac{1}{\mu_\theta(\mu_\theta - 1)}\frac{d\mu_\theta}{d\theta} = \frac{1-\rho_\theta}{\rho_\theta}\frac{d\log\lambda_\theta}{d\theta}.$$

We use

$$d\left[\log(\mu-1) - \log\mu\right] = \left(\frac{1}{\mu-1} - \frac{1}{\mu}\right)d\mu = \frac{d\mu}{\mu(\mu-1)}.$$

We get

$$\log\left(\frac{1-\frac{1}{\mu_\theta}}{1-\frac{1}{\mu_{\theta^*}}}\right) = -\log(\frac{\sigma_\theta}{\sigma_{\theta^*}}) = \int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'.$$

This can be rewritten as

$$1 - \frac{1}{\mu_\theta} = \left(1 - \frac{1}{\mu_{\theta^*}}\right)e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'},$$

$$\sigma_\theta = \sigma_{\theta^*}e^{-\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}.$$

We are targeting

$$1-\frac{1}{\bar{\mu}}=\int_{\theta^*}^{1}\left(1-\frac{1}{\mu_\theta}\right)\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta=\left(1-\frac{1}{\mu_{\theta^*}}\right)\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta.$$

This implies that

$$1-\frac{1}{\mu_{\theta^*}}=\frac{1-\frac{1}{\bar{\mu}}}{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta},$$

$$\sigma_{\theta^*}=\frac{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta}{1-\frac{1}{\bar{\mu}}}.$$

This in turn means that

$$\left(1-\frac{1}{\mu_\theta}\right)=\left(1-\frac{1}{\bar{\mu}}\right)\frac{e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}}{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta},$$

$$\sigma_\theta=\frac{1}{1-\frac{1}{\bar{\mu}}}\frac{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta}{e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}}.$$

These calculations have direct implications for γ_{θ^*}. Indeed, we get

$$\mu_{\theta^*}-1=\frac{\frac{1-\frac{1}{\bar{\mu}}}{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta}}{1-\frac{1-\frac{1}{\bar{\mu}}}{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta}}=\frac{1}{\frac{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta}{1-\frac{1}{\bar{\mu}}}-1},$$

$$\gamma_{\theta^*}=\frac{g(\theta^*)}{1-G(\theta^*)}\frac{1}{\frac{d\log A_\theta}{d\theta}|_{\theta=\theta^*}}=\frac{g(\theta^*)}{1-G(\theta^*)}\frac{\rho_{\theta^*}}{\frac{d\log\lambda_\theta}{d\theta}|_{\theta=\theta^*}}\frac{1}{\mu_{\theta^*}-1},$$

and hence

$$\gamma_{\theta^*}=\frac{g(\theta^*)}{1-G(\theta^*)}\frac{\rho_{\theta^*}}{\frac{d\log\lambda_\theta}{d\theta}|_{\theta=\theta^*}}\left[\frac{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta}{1-\frac{1}{\bar{\mu}}}-1\right],$$

or

$$\gamma_{\theta^*}=\frac{g(\theta^*)}{1-G(\theta^*)}\frac{\rho_{\theta^*}}{\frac{d\log\lambda_\theta}{d\theta}|_{\theta=\theta^*}}\left[\frac{1+\bar{\mu}\left(\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta-1\right)}{\bar{\mu}-1}\right].$$

We now study δ_θ. We have

$$\frac{d\delta_\theta}{d\theta} = (\mu_\theta - \delta_\theta)\frac{d\log\lambda_\theta}{d\theta},$$

$$\frac{d\delta_\theta}{d\theta}\lambda_\theta + \delta_\theta\frac{d\lambda_\theta}{d\theta} = \mu_\theta\frac{d\lambda_\theta}{d\theta},$$

$$\lambda_\theta\delta_\theta = \lambda_{\theta^*}\delta_{\theta^*} + \int_{\theta^*}^{\theta}\mu_{\theta'}\frac{d\lambda_{\theta'}}{d\theta'}d\theta'.$$

We target

$$\bar{\delta} = \int_{\theta^*}^{1}\lambda_\theta\delta_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta = \lambda_{\theta^*}\delta_{\theta^*} + \int_{\theta^*}^{1}\int_{\theta^*}^{\theta}\mu_{\theta'}\frac{d\lambda_{\theta'}}{d\theta'}d\theta'\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta.$$

This implies that

$$\lambda_{\theta^*}\delta_{\theta^*} = \bar{\delta} - \int_{\theta^*}^{1}\int_{\theta^*}^{\theta}\mu_{\theta'}\frac{d\lambda_{\theta'}}{d\theta'}d\theta'\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta,$$

$$\delta_{\theta^*} = \frac{\bar{\delta} - \int_{\theta^*}^{1}\int_{\theta^*}^{\theta}\mu_{\theta'}\frac{d\lambda_{\theta'}}{d\theta'}d\theta'\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta}{\lambda_{\theta^*}} = \frac{\bar{\delta} - \int_{\theta^*}^{1}\int_{\theta^*}^{\theta}\mu_{\theta'}\lambda_{\theta'}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta}{\lambda_{\theta^*}},$$

and hence

$$\bar{\delta}-\delta_{\theta^*} = \frac{\int_{\theta^*}^{1}\int_{\theta^*}^{\theta}\mu_{\theta'}\frac{d\lambda_{\theta'}}{d\theta'}d\theta'\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta - \bar{\delta}(1-\lambda_{\theta^*})}{\lambda_{\theta^*}} = \frac{\int_{\theta^*}^{1}\int_{\theta^*}^{\theta}\mu_{\theta'}\lambda_{\theta'}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta - \bar{\delta}(1-\lambda_{\theta^*})}{\lambda_{\theta^*}}.$$

We want to compute

$$\xi^{\theta^*} = (\mathbb{E}_\lambda[\delta_\theta] - \delta_{\theta^*})\left(\lambda_{\theta^*}\gamma_{\theta^*}\frac{\sigma_{\theta^*} - \mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta]}{\sigma_{\theta^*}-1}\right)\left(\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\right).$$

We get

$$\xi^{\theta^*} = \left(\frac{\int_{\theta^*}^{1}\int_{\theta^*}^{\theta}\mu_{\theta'}\lambda_{\theta'}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta - \bar{\delta}(1-\lambda_{\theta^*})}{\lambda_{\theta^*}}\right)$$
$$\times\left(\lambda_{\theta^*}\frac{g(\theta^*)}{1-G(\theta^*)}\frac{\rho_{\theta^*}}{\frac{d\log\lambda_\theta}{d\theta}|_{\theta=\theta^*}}\left[\frac{1+\bar{\mu}\left(\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta - 1\right)}{\bar{\mu}-1}\right]\frac{\frac{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta}{1-\frac{1}{\bar{\mu}}} - \frac{1}{1-\frac{1}{\bar{\mu}}}}{\frac{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta}{1-\frac{1}{\bar{\mu}}} - 1}\right)$$

$$\times\left(1-\frac{1}{\bar{\mu}}\right),$$

or

$$\xi^{\theta^*}=\left(\int_{\theta^*}^{1}\int_{\theta^*}^{\theta}\mu_{\theta'}\frac{d\lambda_{\theta'}}{d\theta'}d\theta'\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta-\bar{\delta}(1-\lambda_{\theta^*})\right)$$
$$\times\left(\frac{g(\theta^*)}{1-G(\theta^*)}\frac{\rho_{\theta^*}}{\frac{d\log\lambda_\theta}{d\theta}|_{\theta=\theta^*}}\left[\frac{1+\bar{\mu}\left(\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta-1\right)}{\bar{\mu}-1}\right]\right.$$
$$\left.\left(1-\frac{1}{\bar{\mu}-1}\frac{1}{\frac{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta}{1-\frac{1}{\bar{\mu}}}-1}\right)\right)\times\left(1-\frac{1}{\bar{\mu}}\right).$$

We also want to compute

$$\xi^{\epsilon}=(\mathbb{E}_\lambda[\delta_\theta]-1)\left(\mathbb{E}_\lambda[\sigma_\theta]-\mathbb{E}_{\lambda(1-1/\mu)}[\sigma_\theta]\right)\left(\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\right).$$

We get

$$\xi^{\epsilon}=(\bar{\delta}-1)\left(\mathbb{E}_\lambda[\frac{1}{1-\frac{1}{\mu_\theta}}]-\frac{1}{1-\frac{1}{\bar{\mu}}}\right)\left(1-\frac{1}{\bar{\mu}}\right),$$

or

$$\xi^{\epsilon}=(\bar{\delta}-1)\left(\mathbb{E}_\lambda\left[\frac{1-\frac{1}{\bar{\mu}}}{1-\frac{1}{\mu_\theta}}\right]-1\right),$$

or

$$\xi^{\epsilon}=(\bar{\delta}-1)\left(\mathbb{E}_\lambda\left[\frac{1-\frac{1}{\bar{\mu}}}{1-\frac{1}{\mu_\theta}}\right]-1\right),$$

$$\xi^{\epsilon}=(\bar{\delta}-1)\left[\left(\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta\right)\left(\int_{\theta^*}^{1}e^{-\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_\theta\frac{g(\theta)}{1-G(\theta^*)}d\theta\right)-1\right].$$

Finally, we want to compute

$$\xi^{\mu}=\left(\mathbb{E}_\lambda\left[(1-\rho_\theta)\left(1-\frac{\mathbb{E}_\lambda[\delta_\theta]-1}{\mu_\theta-1}\right)\right]\right)\left(\mathbb{E}_\lambda\left[\frac{1}{\sigma_\theta}\right]\right).$$

We have

$$\frac{1}{\mu_\theta - 1} = \frac{1-\left(1-\frac{1}{\bar{\mu}}\right)\frac{e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}}{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_{\theta}\frac{g(\theta)}{1-G(\theta^*)}d\theta}}{\left(1-\frac{1}{\bar{\mu}}\right)\frac{e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}}{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_{\theta}\frac{g(\theta)}{1-G(\theta^*)}d\theta}}$$

$$= \frac{\bar{\mu}-(\bar{\mu}-1)\frac{e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}}{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_{\theta}\frac{g(\theta)}{1-G(\theta^*)}d\theta}}{(\bar{\mu}-1)\frac{e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}}{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_{\theta}\frac{g(\theta)}{1-G(\theta^*)}d\theta}}$$

$$= \frac{\bar{\mu}-(\bar{\mu}-1)\frac{e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}}{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_{\theta}\frac{g(\theta)}{1-G(\theta^*)}d\theta}}{(\bar{\mu}-1)\frac{e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}}{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_{\theta}\frac{g(\theta)}{1-G(\theta^*)}d\theta}},$$

and so

$$\xi^{\mu} = \left(\mathbb{E}_{\lambda}\left[(1-\rho_\theta)\left(1-\frac{\bar{\delta}-1}{\bar{\mu}-1}\frac{\bar{\mu}-(\bar{\mu}-1)\frac{e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}}{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_{\theta}\frac{g(\theta)}{1-G(\theta^*)}d\theta}}{\frac{e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}}{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_{\theta}\frac{g(\theta)}{1-G(\theta^*)}d\theta}}\right)\right]\right)\left(1-\frac{1}{\bar{\mu}}\right),$$

or

$$\xi^{\mu} = \left(\mathbb{E}_{\lambda}\left[(1-\rho_\theta)\left(1-\frac{\bar{\delta}-1}{\bar{\mu}-1}\left(1+(\bar{\mu}-1)\frac{1-\frac{e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}}{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_{\theta}\frac{g(\theta)}{1-G(\theta^*)}d\theta}}{\frac{e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}}{\int_{\theta^*}^{1}e^{\int_{\theta^*}^{\theta}\frac{1-\rho_{\theta'}}{\rho_{\theta'}}\frac{d\log\lambda_{\theta'}}{d\theta'}d\theta'}\lambda_{\theta}\frac{g(\theta)}{1-G(\theta^*)}d\theta}}\right)\right)\right]\right)\left(1-\frac{1}{\bar{\mu}}\right).$$